Computers and Chemical Engineering 152 (2021) 107402

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Automatically optimizing dynamic synchronization of individual industrial process variables for statistical modelling

Tim Offermans[a], Ewa Szymańska[b], Geert H. van Kolllenburg[a], Lutgarde M.C. Buydens[a], Jeroen J. Jansen[a,*]

[a] Radboud University, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands
[b] FrieslandCampina, Amersfoort, The Netherlands

ARTICLE INFO

Article history:
Received 18 January 2021
Revised 28 May 2021
Accepted 2 June 2021
Available online 7 June 2021

Keywords:
Dynamic synchronization
Process control
Automation
Optimization
Industry
Statistical modelling

ABSTRACT

Statistical modelling of industrial production data can lead to improved understanding of the process to benefit process monitoring and control routines. The production data required for such models need however to be synchronized in time, a topic sparsely covered in literature. We propose a strategy for data-driven automated optimization of dynamic synchronization of industrial production data, that optimizes the synchronization *per* process variable and can be applied for on-line monitoring in real-time. The strategy is tested and validated for two relevant production facilities, each of which has multiple production lines or configurations. For all lines and configurations, models predicting the production quality from process variables improved in accuracy using the presented per-variable optimization strategy. Although the prediction accuracy for two models would still be insufficient for real-time monitoring and control, process operators and engineers may still obtain novel process understanding from applying the presented strategy on these models.

1. Introduction

Industrial (bio)chemical production facilities have to be carefully monitored and controlled to guarantee consistent turnover of high-quality product that meets customer wishes. A prerequisite for designing accurate control strategies is to understand how changes in the physical state of the plant and process affect quality and other Key Performance Indicators (KPI) of the production. Multivariate latent variable-based methods, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), are commonly used to extract valuable process-specific knowledge from historical production data (Kourti and MacGregor, 1995). Such methods statistically model and identify relationships between physical process variables, such as temperatures, pressures and flow rates, and the production quality of the plant.

The information obtained from a multivariate regression model may complement process understanding obtained from an engineering point of view, as it represents the actual operation of the plant closer than the intended operation as designed. The use of these models is however not limited to analysing historical data only. After calibration, they can also be used to monitor the modelled relationships in *real-time*. In cases where the product quality is costly or difficult to measure frequently, they can for instance be used as a soft-sensor to predict that product quality from process measurements that are readily available at high frequency (Lin et al., 2007).

Production data is often collected asynchronously, due to sensors operating at different measurement intervals and frequencies. However, for the data to be modelled by a multivariate regression method, or any bilinear method, it needs to be synchronized, regardless of whether it is historical or collected in real-time. Measurements need to be available for all modelled process and/or quality variables at the same production times to be able to estimate the relationships between them for those times. Fig. 1A-B illustrate the problem of asynchronously collected data when attempting to regress a product quality variable (Y) on several process variables (X).

Much research has been done on the statistical analysis of industrial production data, and different review articles are available that elucidate on all the different steps required to prepare the data for statistical modelling. These articles discuss for instance variable filtering, missing value imputation, outlier re-

[*] Corresponding author.
E-mail address: jj.jansen@science.ru.nl (J.J. Jansen).

https://doi.org/10.1016/j.compchemeng.2021.107402

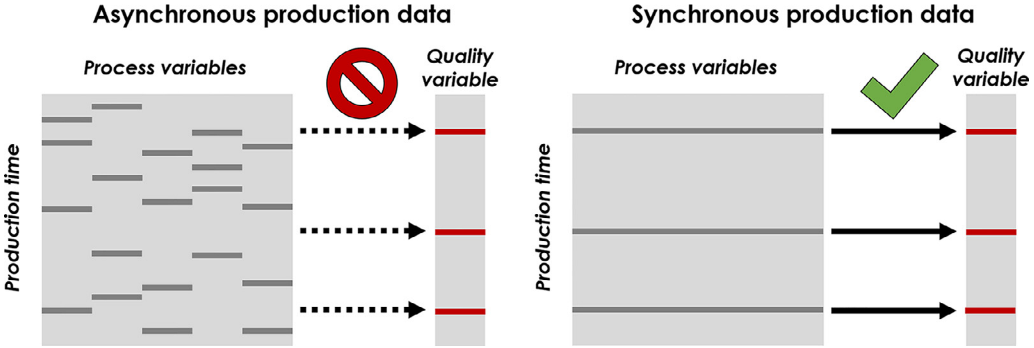

Fig. 1A-B. Visualizations of asynchronous (**A**) and synchronous (**B**) production data. Only for the synchronous data, the relationships between the quality variable and the process variables can be estimated for the production times that the quality variable is available.

moval, nonlinear modelling, dynamic modelling and model validation (Lin et al., 2007; Petr Kadlec and Strandt, 2009; Slišković et al., 2011; Camacho et al., 2008). The issue of dynamic synchronization is however covered limitedly in literature. We have addressed this in a recent publication, where we also demonstrate for an example case study that suffers from data asynchronicity how the choice in synchronization method greatly affects the accuracy of a statistical model (Offermans et al., 2020). Although this work was thorough, two aspects regarding dynamic synchronization of production data were not covered.

Firstly, it was only attempted to find the best *global* synchronization method. The methods tested were taken or adapted from related fields, and include linear interpolation, nearest value interpolation and median-filtering using moving windows. These methods were only applied *globally* to *all* process variables, in the sense that either *all* process variables were synchronized using linear interpolation, or using median-filtering, or using any of the other methods. However, the optimal synchronization method may differ *per* process variable, depending on the sampling rate and dynamics of the variables (which causes the data asynchronicity in the first place). Therefore, to optimally synchronize production data, the best synchronization method would need to be identified *per* process variable.

Secondly, the work offered a critical review of different synchronization methods, but did not offer a protocol or strategy for automated optimization of dynamic synchronization of any given dataset that suffers from asynchronicity. Implementing such a strategy in the default data analysis routine at a production plant would allow process operators to extract more process-specific information from historical data. It is also a valuable additional step in the (re)calibration routine of a statistical model that is used for production monitoring, such as for instance a soft-sensor. This is especially relevant as sensor maintenance and replacement may change the optimal settings for dynamic synchronization over time.

In this work, we propose a strategy for the data-driven automated optimization of dynamic synchronization of process variables for statistical modelling. This strategy not only performs a *global* optimization for all variables, but also a *local* optimization for each individual process variable. This strategy is developed for optimizing production data for a statistical model where a product quality variable is regressed on process variables, and thus optimizes the extraction of statistical relationships between the production process and the production quality. The optimization criterion for the models is the Pearson correlation coefficient between true product quality and product quality as predicted by the model, penalized on data exclusion. A high value signifies an informative model that relates the production quality and process variables well for the majority of the collected data. The strategy will be demonstrated on data from two production facilities in the dairy industry. Both facilities feature multiple production lines or configurations that are independently tested and compared.

2. Methods

2.1. Dynamic synchronization optimization strategy

The proposed strategy for dynamic synchronization optimization of production data for statistical modelling is schematically shown in Fig. 2. The strategy can be divided into three steps. In the first step, the best synchronization method when applied *globally* to all process variables is identified. The second step finds the best synchronization *locally*, for each individual process variable. In the final step, the entire model and method including dynamic synchronization optimization is validated. Each of these three steps as well, as the actual synchronization methods considered, will be explained in detail in the remainder of this Methods-section. The two demonstrator processes on which the proposed strategy is tested are also shortly introduced.

2.2. Synchronization methods

The dynamic synchronization methods considered for each variable are linear, cubic spline, previous value and nearest value interpolation, and window-filtering using means or medians with different window placements and widths. These methods are the same ones as introduced in our earlier work (Offermans et al., 2020), and are exemplified in Figures 3A-H.

For both mean- and median-filtering, the window width and window placement have to be optimized. The width of the window effectively determines the degree of smoothing that is applied to the data, and thus the robustness of the model against outlying process measurements. Ten different window widths are considered for the proposed strategy, evenly ranging from five minutes to five hours. These boundaries were selected so that the average throughput processing times of most chemical production plants, including the ones used for demonstration in this study, fall well within them. The boundaries can however be adapted if the strategy were to be used for a process with a particular long or short processing time.

Four options for window placement are considered: either 100%, 90%, 75% or 50% of the window is placed before the target production time of the synchronization. These four placements are shown in Fig. 1 Fig. 3E-H. Effectively, each cross-combination of either mean- or median-filtering with all ten window widths and with all four window placements is considered as a separate synchronization method. This brings the total number of synchronization methods tested to 84: 4 interpolation methods and 80 (2 × 10 × 4) window-filtering methods.

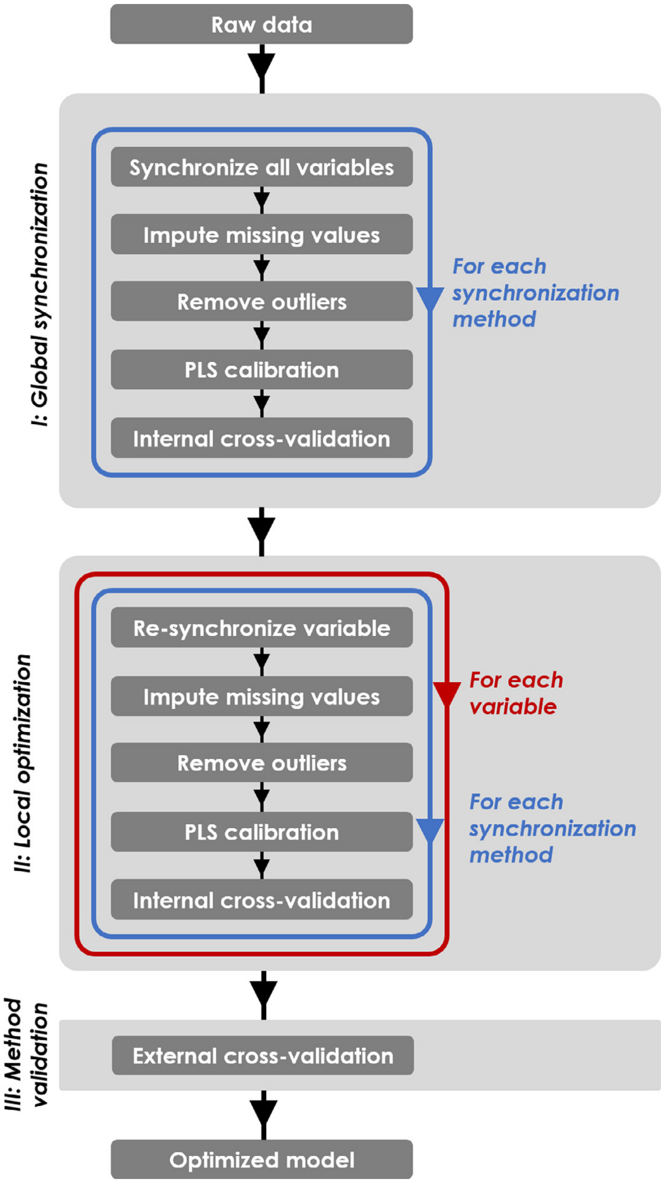

Fig. 2. Schematic representation of the strategy for dynamic synchronization optimization of process variables.

2.3. Step I: global optimization

For the global synchronization optimization, each of the synchronization methods is applied to all process variables universally, yielding a total of 84 synchronized datasets. These datasets are first cleared of missing and outlying measurements, for which the procedures will be discussed in the upcoming subsections. Then, the synchronized and cleaned datasets are statistically modelled by regressing the critical product quality on the (synchronized) process variables, using Partial Least Squares (PLS) regression (Geladi and Kowalski, 1986). Each dataset is mean-centred and autoscaled prior to modelling, as process variables are measured in different units (Gurden et al., 2001).

The accuracy of the models, and thus the reliability of the information given by them, is quantified in terms of the Pearson correlation coefficient r between predicted and reference product quality. A high value of r signifies an informative model that can relate the production quality well to the process variables. The models are subjected to double cross-validation to ensure that the r reflects the accuracy of independent testing data. The inner validation loop is used to select the optimal number of latent variables for that model, and the outer validation loop is used to test the model's accuracy given that number of latent variables, as is proposed in (Szymańska et al., 2012). Both loops used a 5-fold Venetian blinds resampling scheme. This scheme was selected as it ensures that the entire production period that is modelled is equally well represented in the test and training set of each validation fold. Note that this validation is carried out internally in the second step of the strategy, and differs from the additional layer of validation in the third step of the strategy, as will be further explained later on.

The goal is to maximize the model accuracy, and as such the dynamic synchronization method that leads to the PLS model with the highest validated r is selected as the global optimal method. Because the goal is to maximize model accuracy, and for conciseness, comparing different synchronization methods and the models they yield through significance testing with for instance CV-ANOVA is not further discussed (Indahl and Næs, 1998). Before selecting the model with the highest accuracy, the accuracy measures are penalized on data exclusion. The number of data points that are successfully synchronized by each synchronization method, and that are not missing or outlying, can be different. This is elaborately discussed in our previous publication on data synchronization, where the fraction of retained samples ranged from around 0.15 to 0.85 for the different methods studied. Especially window-filtering using windows that are relatively small with respect to the sampling frequency can lead to very few samples available for modelling, as will be explained in more detail Section 2.4.

To prevent the strategy from selecting a synchronization method that leads to a model that is highly accurate but only on a small portion of the data as optimal, the fraction of data points that were successfully synchronized is calculated. This fraction is multiplied with the r for each corresponding PLS model, and the synchronization method leading to the highest value for this measure is selected as optimal. In preliminary studies, penalizing on the squared fraction of data points present was investigated, as this would penalize synchronization methods leading to very few samples relatively more. This however led to very comparable results, as the main goal of excluding synchronization methods leading to very few samples is achieved regardless of whether the penalty is transformed or not.

2.4. Missing values

Missing values may be present in each of the datasets synchronized with one of the window-filtering methods. When synchronizing by calculating either a mean or a median over a moving window, it can and does occur that no value is available for one or more variables within the window at a certain point in time. In these cases, no mean or median can be calculated and matched to the target process quality value, and a missing value is synchronized instead. The locations of missing data thus depend on the window size and placement: especially synchronizations with small window widths are more likely to introduce missing values. The locations of missing data will therefore differ per synchronization method.

Missing values have to be either cleared or imputed from a synchronized dataset before the dataset can be modelled using PLS (Walczak and Massart, 2001). For the presented strategy, missing values are imputed by replacing them with the median value of the data, an approach that is also suggested for modelling industrial production data by Suoza et al. (2016). This imputation is done per synchronized dataset and per process variable.

There are other, arguably more advanced methods available for missing data imputation for PLS modelling (Petr Kadlec and

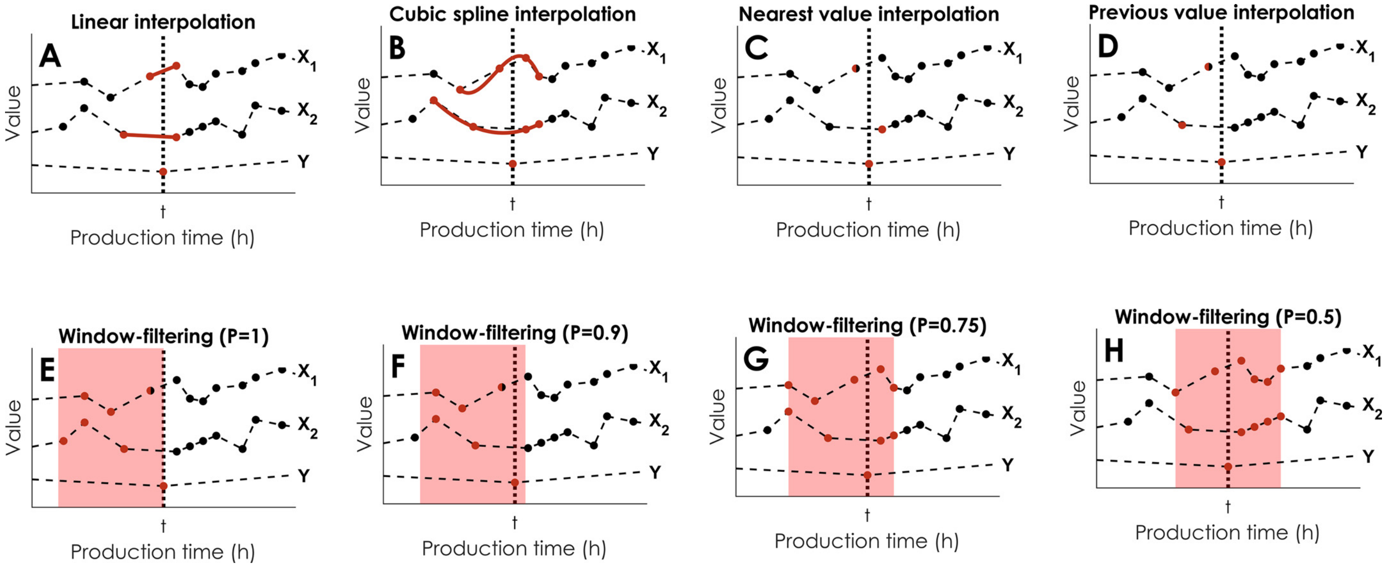

Fig. 3A-H. The different synchronization methods considered for the optimization strategy, examplified using dummy data. Process variables X_1 and X_2 are synchronized to product quality variable Y at production time t using linear (**A**), cubic spline (**B**), nearest value (**C**) or previous value (**D**) interpolation, or using window filtering (**E-H**). For window-filtering, which can be done using means and medians and for variable window width, the four window placements P considered in this study are shown (**E-H**).

Strandt, 2009; Walczak and Massart, 2001; Fortuna et al., 2005; Arteaga and Ferrer, 2002). These methods are however typically slower to calculate and require additional parameters to be optimized. As this would increase the synchronization optimization time and complexity, the (less complex) method of substituting missing values by medians was chosen. Replacing the missing values by means would be more accurate, since the data is mean-centred before PLS modelling. Substituting a missing value by the variable mean would then correspond to setting the contribution of that variable for that sample to zero. However, imputing using medians offers greater robustness against outliers, which are not yet filtered during the imputation but are during the PLS modelling (and mean centring).

2.5. Outliers

Outlying values for the process variables may be manifested in the synchronized data for different reasons, including system errors, production errors or because the data corresponds to non-effective production periods such as pauses, cleaning or breakdowns, and are common in industrial datasets (Wang et al., 2010). As these values do not reflect effective production time of the plant, they have to be removed from the data to optimize the accuracy of the model. Most of the non-physical data is automatically removed during data synchronization, as the process variables are synchronized to the product quality that is sampled only when the plant is in effective production anyway. Some outliers may however remain after synchronization, and are therefore detected and removed. This is done per synchronized dataset individually, as the manifestation of the outliers will be different in each of them.

The outliers are identified using the Hotelling T^2 and Q-statistic, which are calculated from Principal Component Analysis (PCA) models (Varmuza and Filzmoser, 2009). These models are calculated for each dataset, autoscaling the data beforehand and using as many principal components are required to describe at minimal 70% of variance in the dataset. Any sample for which either the Hotelling T^2 or the Q-statistic (or both) are over two standard deviations removed from the median value, is marked as outlier and is removed from the respective dataset (Lin et al., 2007). An (additional) univariate outlier removal step on the individual unsynchronized process variables was considered, but not included as the unsynchronized process variables still contain much data corresponding to non-effective production periods such as cleaning. These periods impact the (automatic) estimation of the variable median and standard deviation, which reduces the stability and accuracy of the optimization.

2.6. Step II: local optimization

After the global optimization has been completed, all synchronization methods are re-considered iteratively per variable by order of importance. The measure for importance used is the absolute regression coefficient, which are assigned to each process variables by the PLS algorithm (Geladi and Kowalski, 1986). Other measures for variable importance were considered, such as Variable Importance in Projection (Eriksson et al., 2013), Selectivity Ratio (Rajalahti et al., 2009) and Significance Multivariate Correlation (Tran et al., 2004). The absolute regression vector was chosen as it directly reflects the relationships between the process variables and the product quality (Wang et al., 2015), which is what the strategy is intended to optimize.

The local optimization starts with the variable that has the highest absolute regression coefficient. It re-considers all synchronization methods for only this variable, while keeping the methods selected for the other variables unchanged. This results in 84 (new) synchronized datasets, for which the process variable being optimized is synchronized with any of the 84 synchronization methods as introduced in Section 2.2, and for which all other process variables are identical. All datasets are subjected to the same missing value imputation, outlier removal and PLS modelling and validation steps as used during the global optimization. The synchronization method leading to the PLS model with the highest validated accuracy is chosen as optimal for the variable being optimized. These accuracies are, as with the global optimization, penalized on data exclusion.

The variable that has the highest absolute regression coefficient in that same model (and is not already optimized) is optimized in the next iteration. This is repeated until all variables are optimized. Once the optimal method is found for a certain variable, its selection is fixed and is used instead of the global optimal method when optimizing the next individual variable(s). Note that the regression coefficients, and thus the order of importance, of the yet-to-optimize variables may change after each iteration. The PLS model with the highest validated (penalized) accuracy found during the optimization of the last variable is selected as the final model, with optimal dynamic synchronization for each separate process variable.

Locally optimizing the synchronization method for an individual variable in the presented way will likely improve the overall performance of the model (in case a better synchronization method than the global method is found for that variable), or may leave the overall performance unaffected (in case the global method was already the best method for that variable). The local synchronization will however never decrease the model performance, as re-selecting the global method as the best local method for that variable is the worst-case scenario and will not affect the overall model performance. In general, a higher increase in model performance may be expected for the first few variables that are locally optimized, as they are sequentially optimized in order of decreasing importance.

Asynchronous data always has to be synchronized in some way before it can be modelled, which is why the local optimization cannot be used without the global optimization. During synchronization optimization of one variable, all other non-optimized variables still have to be synchronized. It is technically possible to use for instance linear interpolation as a default method for this, but we chose to use the global optimal method instead. This ensures that the synchronization of the variables that are not being optimized is still to a certain degree optimal. Because of the multivariate nature of the data and the models, this increases the accuracy of the synchronization method selection of the variable that is being optimized.

2.7. Step III: method validation

To ensure that the optimal synchronization settings and associated model accuracy are not overfitting the modelled data, the entire global and local optimization has to be subjected to another layer of (cross-)validation. This cross-validation effectively acts as a complete external and independent third layer of cross-validation, on top of the double cross-validation used to optimized the individual PLS models. A 5-fold Venetian blinds resampling scheme is also used for this validation layer. The modelling performance found after this the cross-validation layer gives an estimate of how well newly measured production data would be modelled using the optimal settings found by the proposed strategy. The reported performances are the average performances found for the five models, one calibrated per validation fold.

Cross-validation is generally recommended for the synchronization optimization strategy and used to demonstrate the strategy. This because cross-validation ensures a validated result that accurately represents the entire production period modelled, also for datasets with a limited sample availability. For datasets for which many samples are available, using a single independent test set for validation is however also likely accurate, and can be considered as it would save calculation time.

The presented approach for synchronization optimization and model calibration is computationally intensive, because of the elaborate validation scheme and because all synchronizations have to be calculated and tested for all process variables. Applying an optimized set of synchronization methods to incoming data and projecting that data into a calibrated prediction model is however not intensive. It should also be taken into account that asynchronously collected data always has to be synchronized with one method or another. The applicability of the presented approach to process monitoring in real-time is therefore little to not limited. Updating the model may take more time than is usual for a soft-sensor without synchronization optimization, but such updating is typically not done frequently enough for the longer calibration time to be limiting.

Although the dynamic synchronization is optimized for the reported models in an advanced way, there are certain possibly relevant aspects that are not optimized. Such aspects include for instance variable selection and nonlinear modelling. These steps were not in scope for this work, but could be considered for future use of the presented strategy, as they can be valuable additions that improve the modelling accuracy further.

2.8. Demonstrator process I: lactose powder production

The first demonstrator process for this study is a facility that produces dry lactose crystal powder from aqueous lactose. The crystals are first grown in a crystallization tank, after which they are centrifuged and subjected to two consecutive drying steps to form the dry powder product. Different parallel instruments are available for all process steps, which are activated in pre-defined configurations depending on consumer, maintenance or cleaning wishes. The critical production quality parameter or KPI for this process is the mass fraction of small crystals in the product (fines). As this mass fraction can currently only be measured off-line a few times per day, improved understanding or even prediction of it from production data would benefit the overall controllability of the plant.

Historical data was collected for a period of 39 months. Only process variables from the centrifuge and drying steps are used as predictor variables, as these steps are believed to be the major sources affecting the fraction of crystal fines. The processing time of these steps is 30–60 min, depending on the capacity that the plant is running on. Only data corresponding to the two most often used preset operation configurations were used. These configurations are henceforth referred to as configurations A and B, and were subjected to the soft-sensor optimization strategy individually. Measurements for 48 equivalent process variables were collected for both configurations. These variables are for instance temperatures, flow rates, power consumptions and pressures, and have average sampling intervals between 15 s and 5 min. The total number of samples collected for configurations A and B are 868 and 912, respectively.

2.9. Demonstrator process II: milk protein powder production

The second process on which the dynamic synchronization optimization strategy is demonstrated is a milk protein powder production facility. This is the same facility as reported in our earlier publication (Offermans et al., 2020). The protein powder is produced from skim milk by heating, precipitation, washing and drying, and the total throughput time of this plant is around 30 min. The critical product quality parameter for this process is the mineral content in the milk protein powder, which should be as low as possible. Like the mass fractions for demonstrator process I, this mineral content can only be measured with off-line laboratory analysis a few times per day. A regression model predicting it from the process variables would therefore benefit the understanding, monitoring and control of the plant.

The plant features three parallel production lines, which were modelled individually and which are referred to as lines A, B and C. Data corresponding to 45 process variables (equivalent for the three lines) were collected alongside the mineral content for the same 39 months as were collected for the lactose powder production demonstration. The average sampling interval of the process variables ranges from 10 s to 5 min; the sampling interval of the mineral content is around 8 h. The number of mineral content samples collected for lines A, B and C are 1256, 728 and 624, respectively.

Table 1
Modelling performances in terms of Pearson correlation coefficient between predicted and reference product quality for demonstrator process I. Results are given for both operation configurations, for both synchronization optimization strategies (global and local) and for both the validation and calibration data.

Operation configuration	Samples	Variables	Global optimization Calibration r(pred, ref)	 Validation r(pred, ref)	Local optimization Calibration r(pred, ref)	 Validation r(pred, ref)
A	868	48	0.79	0.74	0.88	0.81
B	912	48	0.74	0.70	0.87	0.78

3. Results & discussion

In this section, regression models for each of the demonstrator processes and plants as introduced above will be discussed and compared. Modelling accuracies are compared in terms of (validated) Pearson correlation coefficient between modelled and reference product qualities. This measure effectively represents how well the variation in the product quality can be explained from changes in the process itself, and thus how well the model could be used for soft-sensoring. We will furthermore discuss the actual synchronization methods selected for each of the process variables, and compare them within one production plant and between parallel production lines or configurations. This allows us to see if difference in the nature and/or dynamics in the variables indeed call for different synchronization methods. Finally, we will discuss the importance of each the process variables for predicting the product quality, and how those importances change when the synchronization is optimized locally per variable instead of globally for all variables. Studying these importances relatively for a model can lead to a better understanding of which parts of the process are most influential on the production quality, and on how they should be controlled.

3.1. Demonstrator process I: lactose powder production

The accuracies for the regression models calibrated while using only the global synchronization optimization or while also using the local synchronization optimization are given in Table 1, for both operation configurations of demonstrator process I. The validated performance found using the local optimization is higher for both cases, showing that a more accurate model is obtained when the synchronization is optimized per process variable individually.

For all models, the performances on the calibration set is higher than on the validation set. This is expected, as the models will in most cases perform better on seen data than on unseen data. However, the differences between these performances are not so large to suggest that the models are highly overfitting the calibration data. This is especially important for the models found using local optimization. Optimizing the synchronization per variables increases the complexity of the model significantly, which increases the risk over model overfitting. The absence of such overfitting shows that the validation routines used within the synchronization optimization strategy are accurately, and results in a reliable optimization.

Fig. 4A-B show the prediction versus reference plots for the regression models optimized for both configurations. These plots correspond to the validated results found using the local synchronization optimization. For configuration A, there seems to be little to no samples with clear outlying prediction accuracies. For configuration B however, there are some samples with outlying accuracy, each of which has a very low reference value. Because of their low reference mass fraction values, it is likely that these samples suffer from sampling, analysis or registration errors for those measurements. This is affirmed by them not being removed by the outlier removal procedure, which only determines outliers based on the independent process data (X) and not on the dependant mass fraction data (Y). That outlier removal procedure was selected because the goal of the strategy is to synchronize the process data to the product quality data, and the choice in synchronization method for the process data does not change the values for the product quality data. The outliers in Figure 4B do however signify that it is essential to carefully remove samples with outlying product quality before employing the proposed synchronization optimization strategy, preferably with process experts knowledge.

The synchronization methods that were selected by the optimization strategy for each variable are given in Table 2, for both operation configurations. The row *'All'* refers to the method selected by the global optimization; all subsequent rows refer to the synchronization method for one particular variable using the local optimization step.

The optimal global method for both configurations is to use previous value interpolation. This would theoretically be the most accurate method as it matches each product quality sample to the process values that are last known and thus most relevant in time. Remarkable is that previous value interpolation in general outperforms any form of window-filtering for this process. Window-filtering would offer a higher robustness against outlying values in the process variables due to a smoothing effect, which suggests that the process variables for this demonstrator process suffer little from outliers. It also suggests that the values for the process variables are changing relatively rapidly over time, and that these result in quick response changes in the product quality. This implies that the system has a high responsiveness in general, signifying the need for fast control action formulation and thus for a model (soft-sensor) predicting the product quality in real-time. However, it should also be noted that missing value interpolation cannot lead to missing values while window-filtering can, at least for the implementation in the presented strategy. As the synchronization methods are optimized towards both high modelling accuracy and minimum number of missing values, nearest value interpolation has an added advantage over window-filtering.

There is high diversity in the synchronization methods found optimal for the individual variables. Linear, cubic spline, previous and nearest value interpolation are selected most often. In comparison to window-filtering, all these methods use only data measured close to the production quality in time. As such, these variables likely suffer little from outliers and change frequently. The variables for which window-filtering is found optimal are likely more prone to outliers, which is supported by median-filtering being selected more often than mean-filtering. These variables may also change more gradually and slower over time, and cause more long-term responses in the product quality. This is confirmed by the fact that if window-filtering is chosen, long windows in comparison to the total throughput time of the plant are selected. No relationship could be found between the physical property measured (level, flow, temperature, etc.) and the synchronization method chosen.

The diversity in the methods chosen per process variables shows that the dynamics of these variables and the responsiveness of the product quality to changes in these variables are quite different, and that the synchronization method should indeed be opti-

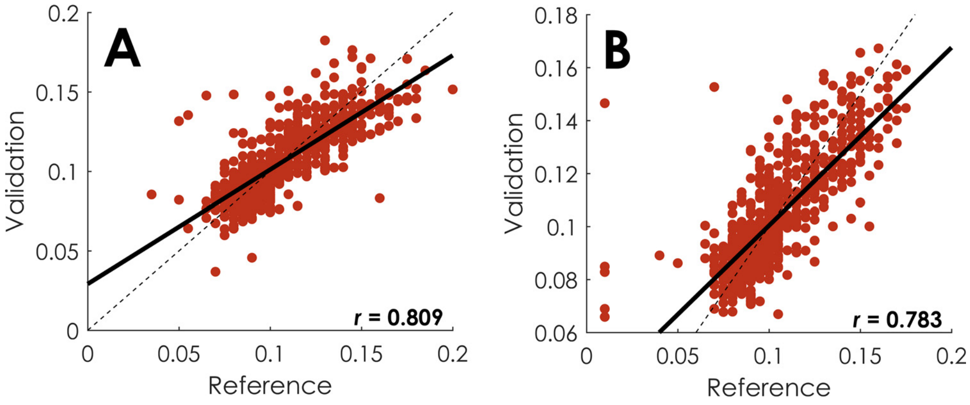

Fig. 4A-B. Prediction versus reference plots for each operation configuration (A and B) of demonstrator process II. These results were found after optimizing the synchronization using the local optimization method and corresponding to the cross-validated results.

Table 2
Synchronization methods found optimal for all variables (using global optimization) and for each variable (using local optimization), for both operation configurations of demonstrator process I. The names consists of the interpolation method or, in case of window-filtering, to the population estimator followed by the window width in minutes and the window placement as illustrated in Figures 3e-h.

Process variable	Operation configuration A	Operation configuration B	Process variable	Operation configuration A	Operation configuration B
All	previous	previous	25	previous	mean-300–0.5
1	nearest	nearest	26	median-71–0.5	median-234–1
2	spline	spline	27	previous	linear
3	mean-300–0.5	previous	28	mean-267–0.5	previous
4	nearest	spline	29	previous	linear
5	previous	previous	30	nearest	nearest
6	linear	linear	31	previous	nearest
7	nearest	nearest	32	linear	spline
8	linear	previous	33	nearest	mean-267–0.5
9	linear	median-202–1	34	mean-300–1	previous
10	nearest	nearest	35	previous	nearest
11	linear	median-136–0.75	36	previous	median-300–1
12	spline	median-136–0.5	37	median-267–0.5	median-103–0.5
13	linear	nearest	38	mean-267–0.75	previous
14	spline	median-169–0.5	39	nearest	previous
15	linear	spline	40	previous	previous
16	median-38–0.5	nearest	41	mean-136–0.9	linear
17	mean-169–0.5	linear	42	nearest	nearest
18	linear	linear	43	median-71–0.5	median-38–0.5
19	median-300–1	linear	44	nearest	nearest
20	nearest	nearest	45	median-136–1	median-38–0.5
21	linear	previous	46	previous	previous
22	nearest	nearest	47	mean-169–0.5	previous
23	nearest	previous	48	median-136–0.9	median-169–0.75
24	linear	nearest			

mized *per* variable. The choice in optimal synchronization method also differs between the two operation configurations. This could be an indication that the strategy is overfitting the synchronization choices per configuration. However, as discussed before, the small differences between the validation and calibration accuracies indicate that the models do not suffer from such strong overfitting. The differences between the choices per configuration do signify the need to model each configuration individually.

To illustrate this further, the data for configuration A was synchronization with the methods found using local optimization for configuration B, and vice versa. All other modelling steps, including validation, were retained to ensure comparability of the results. The modelling results found after external cross-validation are given in Table 3. These results show indeed that for both configurations, the most explanatory models are obtained when they are optimized on the data from that same configuration, as may be expected. However, interchanging the synchronization methods between configurations still give a relative high modelling accuracy for both configurations. This is likely due to the physical comparability of the configurations, and shows that the synchronization optimization does not overfit the configurations.

Table 3
Cross-validated modelling accuracies found for each operation configuration when applying the synchronization methods found after local optimization for the other configuration, for demonstrator process I. All results are given in terms of Pearson correlation coefficient (r) between cross-validated and reference product quality. The results on the main diagonal correspond to those in the rightmost column of Table 1.

		Settings from configuration	
		A	B
Applied to configuration	**A**	0.81	0.77
	B	0.75	0.78

Fig. 5A-B show for both configurations the absolute regression coefficient of each process variable when the model is optimized using the global or the local synchronization optimization. The val-

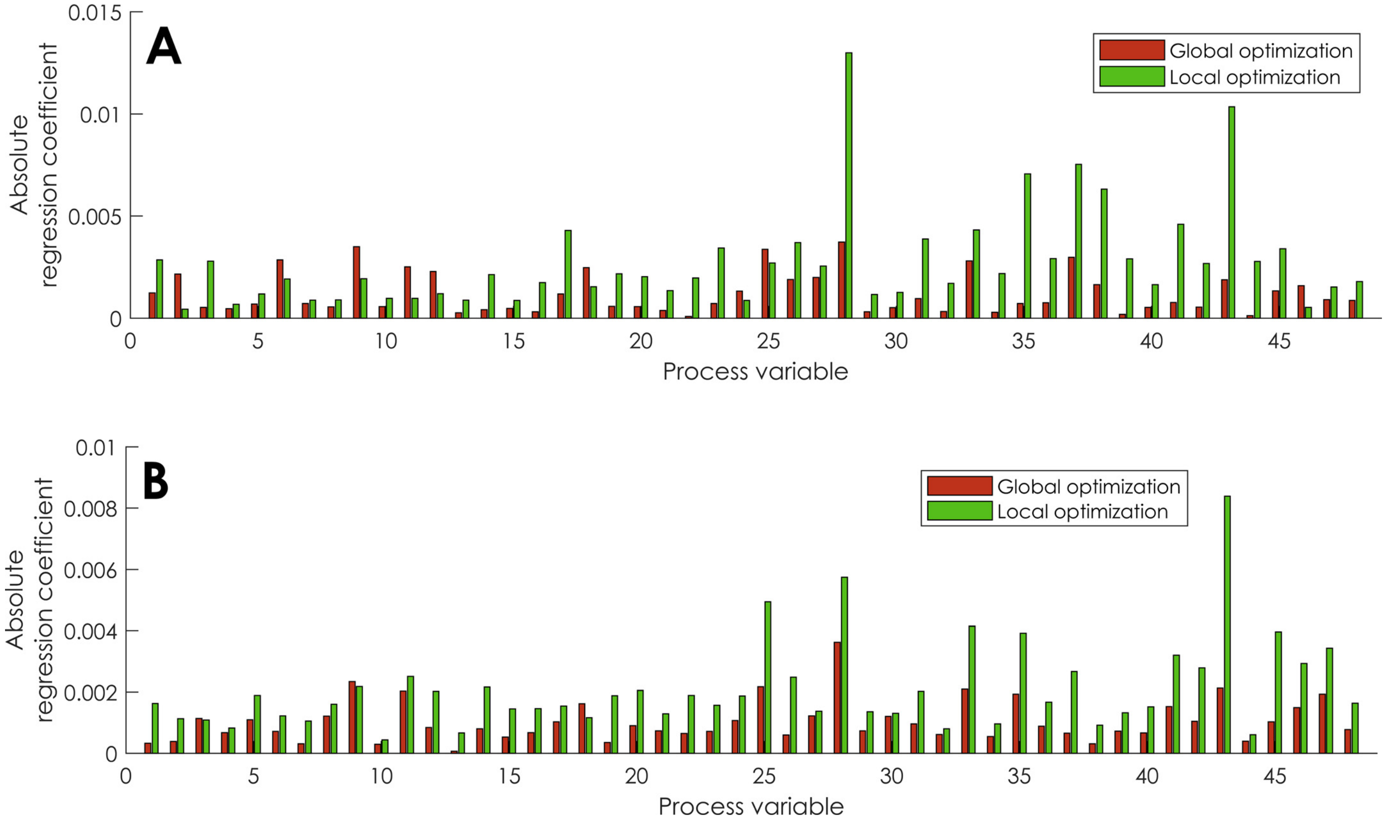

Fig. 5A-B. Absolute regression coefficients found for the models calibrated for demonstrator process I. Results are given for both operation configurations (A and B) and for both using the global and local synchronization during calibration.

ues are averaged over all five validation folds. The absolute regression coefficient of a process variable quantifies the contribution that variable has to the prediction of the product quality, and may be interpreted as a measure of variable importance.

The local optimization optimizes the synchronization method of each individual process variable towards maximum contribution of that variable to the prediction of the product quality. It can therefore be expected that most process variables will overall have a higher absolute regression coefficient after local optimization, as opposed to global optimization. Fig. 5A-B confirm this for most process variables. This holds especially for variables that have a relative high contribution to the globally optimized model, which is sensible as they are optimized first during the local optimization.

Some variables, for instance variables 28, 35, 37, 38 and 43 of configuration A, show an especially high increase in regression coefficient from global to local synchronization. This signifies that process variables are indeed able to contribute more to the prediction of product quality when their synchronization method is optimized individually. For some variables, the absolute regression coefficient decreases when the local optimization is used. This is due to the multivariate nature of the data and the regression models used. Optimizing the synchronization of one process variable can increase the contribution of that variable a lot, but decrease the contribution of a related process variable somewhat also, regardless of the synchronization method used for that related variable. Remarkable is that variables 28 and 43 have a very high contribution to the models of both configurations, despite them having slightly (variable 43) or considerably (variable 28) different optimal synchronization methods for both configurations. This shows that equivalent variables can be important in both configurations, but may require a different synchronization method.

3.2. Demonstrator process II: milk protein powder production

Table 4 shows the accuracies for the regression models calibrated for each production line of demonstrator process II, for both the global and local optimization approach. These results show that also for this demonstrator process, using the local optimization as opposed to the global optimization yields a model with higher validated accuracy.

The validated performance for production line A is high, and no indication of overfitting is present for this model. The performances of the models for production lines B and C on the validation data are however quite low, and the much higher performance on their respective calibration sets does indicate that these models are overfitted. One possible reason for this is that the data for these two production lines contains more noise. Causes for such noise include less stable equipment, more frequent maintenance, higher variation in raw material feed or product demand, or less consistent control practices in general.

Another reason for a higher level of model overfitting is the lower number of samples that are available for these two lines. This is supported by the fact that the model for line C has both the lowest sample count and the lowest performance. As mentioned before, performing the local synchronization optimization adds complexity to the model. To cope with this added complexity and to prevent the model from overfitting, a large number of calibration samples is required. This shows that a large collection of strategically obtained historical data is a prerequisite to reliable calibrate a model using the local synchronization optimization approach.

The models found for production lines B and C have too low performance to use them for real-time process control purposes, despite using the more advanced local synchronization optimiza-

Table 4
Modelling accuracy in terms of Pearson correlation coefficient between predicted and reference product quality for demonstrator process II. Results are given for all production lines, for both synchronization optimization strategies (global and local) and for both the validation and calibration data.

Production line	Samples	Variables	Global optimization Calibration r(pred, ref)	 Validation r(pred, ref)	Local optimization Calibration r(pred, ref)	 Validation r(pred, ref)
A	1256	45	0.75	0.72	0.81	0.75
B	728	45	0.57	0.46	0.73	0.54
C	624	45	0.43	0.34	0.75	0.48

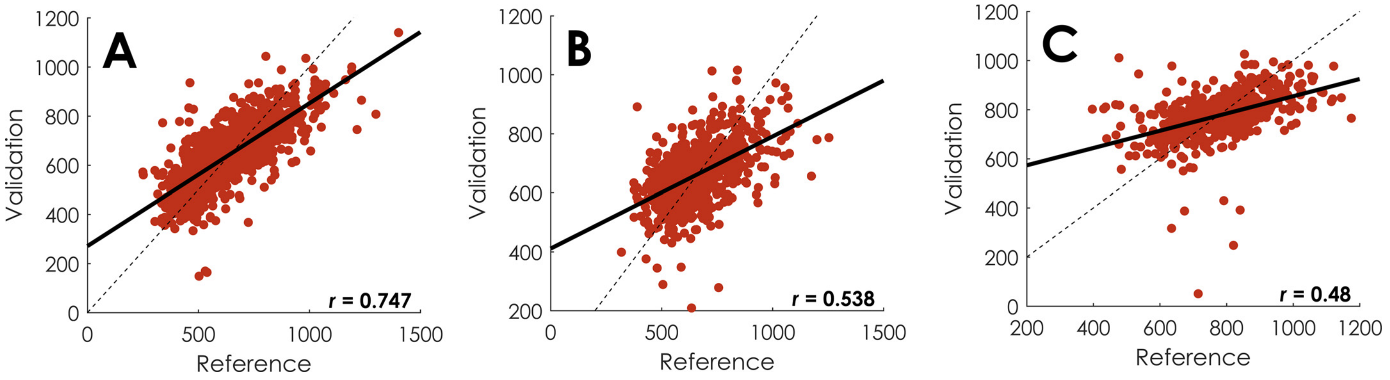

Fig. 6A-C. Prediction versus reference plots for each production line (A to C) for demonstrator process II. These results were found after optimizing the synchronization using the local optimization method and corresponding to the cross-validated results.

tion. However, local synchronization still leads to a better description of the correlations between process variables and product quality. As such, the local optimization approach can still help process operators and engineers to obtain a better understanding of the plant. This may improve monitoring and control practices, and thus higher production quality.

The prediction versus reference plots for the regression models optimized for all production lines are shown in Fig. 6A-C. These figures only show the validated results found using the local synchronization optimization. For production lines B and C there are some samples for with the respective model performs particularly bad. These are mostly samples for which the predicted values are far below the reference values. The presence of these samples is an additional cause for the lower performance of the models for lines B and C. This is confirmed by the model for line C having both having the lowest performance and suffering seemingly most from outliers.

The predicted values for these samples are more outlying than the reference values, which indicates that these inaccurate predictions result from outliers in the process measurements and not from inaccurate product quality measurements. Increasing the sensitivity of the outlier detection method used in the optimization approach may therefore improve the accuracy and reliability of the final model. This also shows that the optimal setting for this outlier detection may be different per production process.

Table 5 shows the synchronization methods found optimal for each process variable and for each production line of demonstrator process II (analogues to Table 2 for demonstrator process II). The globally optimal synchronization method is quite comparable for the three production lines. For all lines, using a median filter that is placed either for 100% or 90% before the target time is optimal. There is some variation in the optimal window width, but they are all wide with respect to the total process throughput time of 30 min. This indicates that changes in the process state can still affect the production quality for a prolonged time. These results are in agreement with our earlier findings for this production facility (Offermans et al., 2020). As for demonstrator process I, the different physical properties measured did not show any clear preference for a certain synchronization method.

From all synchronization methods considered, median filtering would offer the highest robustness against outliers in the process data, especially when relatively wide windows are used. It being selected as best global method for this process suggests that this process suffers from such outliers, and more so than demonstrator process I. This is confirmed by the analysis of the prediction versus reference plots for both processes (Figures 4A-B and 6A-C), and by the fact that relatively wide windows are selected.

As for demonstrator process I, there is quite some variation in synchronization methods found optimal for the individual variables for demonstrator process II. As discussed before, this results from the process variables representing different instrument and measurements with different dynamic behaviour, and signifies the need to optimize the synchronization method per process variable. The choice in optimal methods per variable differs also per production line.

One reason for this is that the production lines are not exact copies from one another, either by design or introduced by maintenance and repair practices. However, the differences between validation and calibration performances were quite high for production lines B and C. This indicates that these models could be overfitted at least to some degree, which can be an alternative reason for the large differences in selected synchronizations between the production lines.

Table 6 shows the validated modelling results obtained for each production line when the optimal synchronization methods of another line are used, and is analogous to Table 3 for demonstrator process I. As with demonstrator process I, the highest accuracies are obtained when optimizing the synchronization on the data from the same line as for which the model is desired, as expected. However, using the optimized synchronization from another production line or configuration does seem to generally lower the modelling accuracy more than was the case for demonstrator process I (save for when using the optimal methods from line C while modelling line A). This could be the result of the lines of demon-

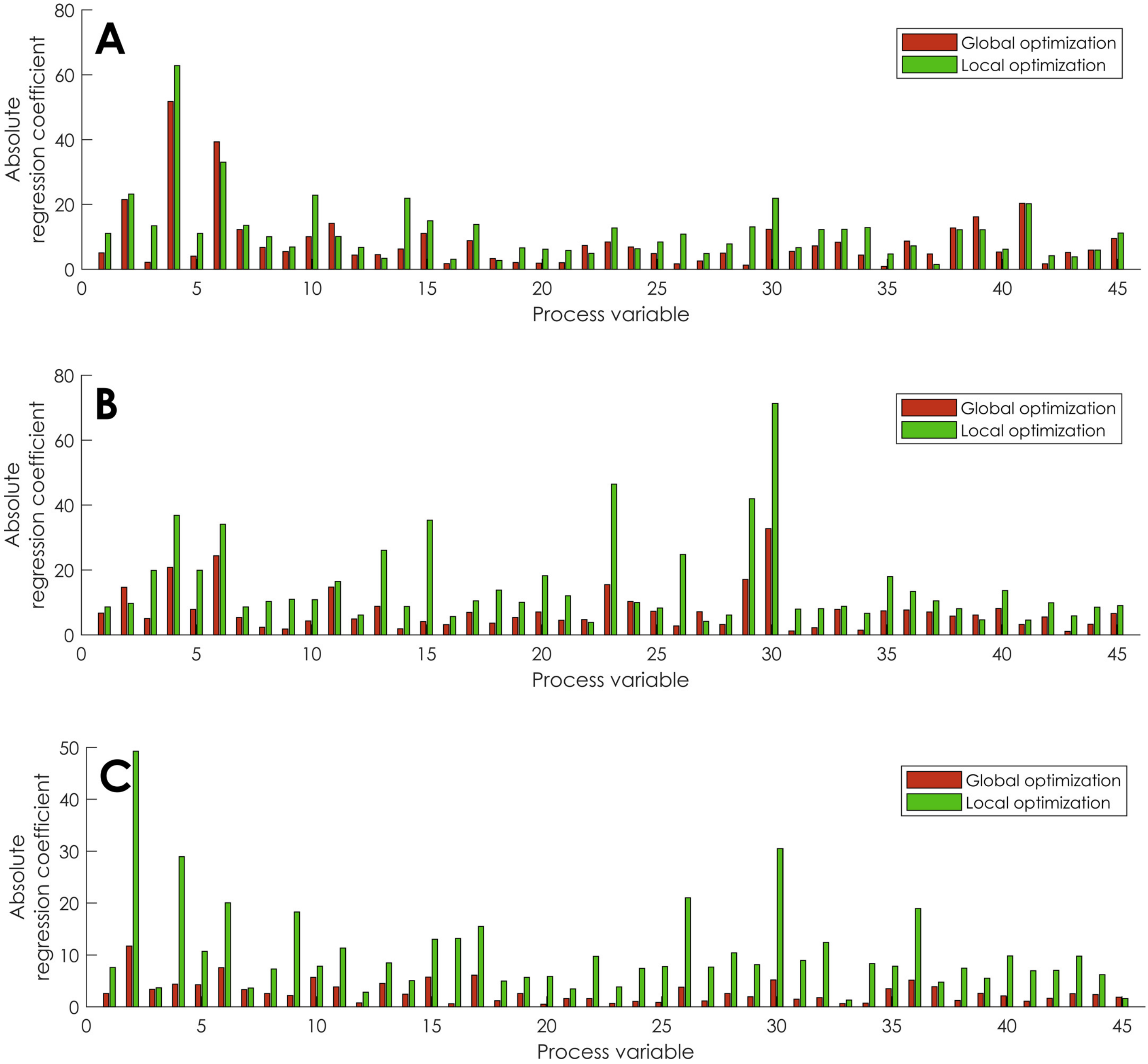

Fig. 7A-C. Absolute regression coefficients found for the models calibrated for demonstrator process II. Results are given for all production lines (A to C) and for both using the global and local synchronization during calibration.

strator process II being physically less comparable than the configurations of demonstrator process I.

The absolute regression coefficients for the models are shown in Fig. 7A-C for all three production lines and for both the global and local optimization approaches. As for demonstrator process I, the absolute regression coefficient is higher when the local optimization is used, especially for variables that have a high coefficient for the globally optimized model. Variable 6 in production line A is an exception. As this variable is the second most important, it is likely correlated to the most important variable: variable 4. Optimizing the synchronization for variable 4 increased its contribution to the prediction of the product quality, but simultaneously decreased the contribution of variable 6 regardless of its synchronization method.

The relative contributions of the process variables to the prediction of the product quality differs for the production lines, and more so than they differed for the two operation configurations for demonstrator process I. This is in line with the higher variation in synchronization method found optimal and the higher variation for the prediction accuracies for the different lines of demonstrator process II. Careful investigation of the results as presented in this discussion can however help process operators and engineers

Table 5
Synchronization methods found optimal for all variables (using global optimization) and for each variable (using local optimization), for each of the production lines of demonstrator process II. The names consists of the interpolation method or, in case of window-filtering, to the population estimator followed by the window width in minutes and the window placement as illustrated in Figures 3e-h.

Process variable	Production line 1	Production line 2	Production line 3
All	median-202–1	median-169–0.9	median-267–1
1	median-169–0.5	nearest	median-38–1
2	mean-5–0.9	previous	mean-267–1
3	spline	mean-234–0.9	mean-300–1
4	mean-169–1	mean-234–1	mean-202–1
5	median-234–0.75	median-38–0.75	median-300–1
6	mean-169–1	median-234–1	mean-136–1
7	median-267–0.9	median-169–0.9	nearest
8	mean-234–1	median-136–0.5	median-38–0.9
9	median-300–1	median-267–1	median-267–1
10	median-169–0.5	mean-169–0.5	median-300–1
11	median-267–1	previous	median-300–1
12	mean-103–0.5	previous	mean-300–1
13	mean-38–1	mean-103–1	median-202–0.75
14	nearest	nearest	previous
15	median-300–1	median-202–1	median-300–1
16	median-38–1	previous	median-300–1
17	mean-169–1	spline	mean-267–1
18	median-103–0.9	median-169–1	median-169–1
19	previous	spline	median-267–1
20	median-169–1	previous	median-300–0.9
21	spline	median-267–1	spline
22	spline	median-267–1	median-202–1
23	median-169–1	median-38–0.75	mean-300–1
24	median-38–0.9	nearest	median-267–1
25	median-169–0.5	linear	median-169–0.5
26	nearest	median-300–0.5	mean-300–0.5
27	spline	spline	median-136–1
28	median-38–0.75	spline	median-103–1
29	median-267–0.75	median-103–1	mean-5–1
30	median-71–1	median-169–0.9	median-300–1
31	median-71–0.75	mean-38–0.5	median-136–0.5
32	median-300–0.75	spline	median-38–0.5
33	median-234–1	nearest	linear
34	median-300–0.75	previous	spline
35	mean-267–1	median-136–1	spline
36	spline	previous	linear
37	spline	spline	spline
38	median-103–0.5	spline	median-136–1
39	median-267–0.75	spline	median-300–1
40	median-234–0.75	spline	mean-300–1
41	median-300–0.9	linear	mean-300–1
42	spline	previous	mean-300–1
43	spline	nearest	spline
44	median-267–0.9	previous	median-71–0.5
45	median-300–0.5	previous	median-136–1

Table 6
Cross-validated modelling accuracies found for each production line when applying the synchronization methods found after local optimization for another line, for demonstrator process II. All results are given in terms of Pearson correlation coefficient (r) between cross-validated and reference product quality. The results on the main diagonal correspond to those in the right-most column of Table 4.

		Settings from line		
		A	B	C
Applied to line	A	0.75	0.68	0.73
	B	0.49	0.54	0.44
	C	0.38	0.36	0.48

to better understand each of the production lines and their differences.

4. Conclusion

In our study, we have developed a new strategy for automatically optimizing the dynamic synchronization of individual process variables for statistically modelling industrial production data. Although the method is specifically designed and tested for regression models that predict the production quality from process variables, it could be extended to models of a different nature. The strategy first optimizes the synchronization globally by finding the method that leads to the most accurate prediction of product quality when applied to all variables universally. It then optimizes the synchronization locally, by iteratively re-considering all synchronization methods for each variable individually. This all while taking into account missing data imputation and outlier removal. To demonstrate the strategy, prediction models were calibrated for data from two demonstrator processes, each for which multiple production configuration or lines were present and modelled separately. All models were calibrated to predict the production quality from process variables, and were cross-validated and elaborately compared. For all models, the local optimization resulted in more accurate predictions than the global optimization did, showing that the more advanced local optimization is a valuable addition when modelling production data suffering from asynchronicity. The choice in optimal synchronization method was found to be dependant on the process, on the production line or configuration, and on the process variable. This variation results from differences in dynamics and the manifestation of outliers for different plants and process variables, and signifies the need to model production lines individually. For three out of five models, the optimized models have high enough accuracy to consider them as soft-sensors for real-time process monitoring and maybe even control. For two models, the presented strategy for per-variable synchronization optimization did improve the accuracies, but the improved accuracy was still too low for the sensors to be used for real-time process monitoring purposes. However, as the optimization strategy still maximized the correlation between the process variables and the end product quality, investigation of these optimized models can lead to an unprecedented understanding of the production process.

Author statement

Tim Offermans: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization

Ewa Szymańska: Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration

Geert H. van Kollenburg: Conceptualization, Methodology, Supervision

Lutgarde M. C. Buydens: Supervision, Project administration, Funding acquisition

Jeroen J. Jansen: Conceptualization, Methodology, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is co-funded by TKI-E&I with the supplementary grant 'TKI- Toeslag' for Topconsortia for Knowledge and Innovation (TKI's) of the Ministry of Economic Affairs and Climate Policy. The authors thank all partners within the project 'Integrating Sensor Based Process Monitoring and Advanced Process Control (INSPEC)', managed by the Institute for Sustainable Process Technology (ISPT) in Amersfoort, The Netherlands.

References

Kourti, T., MacGregor, J.F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemom. Intell. Lab. Syst. 28, 3–21.

Lin, B., et al., 2007. A systematic approach for soft sensor development. Comput. Chem. Eng. 31, 419–425.

Petr Kadlec, B.G., Strandt, S., 2009. Data-driven Soft Sensors in the process industry. Comput. Chem. Eng. 33, 795–814.

Slišković, D., Grbić, R., Hocenski, Ž., 2011. Methods for plant data-based process modeling in soft-sensor development. Automatika 52 (4), 306–318.

Camacho, J., Picó, J., Ferrer, A., 2008. Bilinear modelling of batch process. Part II: a comparison of PLS soft-sensors. J. Chemom. 22, 533–547.

Offermans, T., et al., 2020. Synchronizing process variables in time for industrial process monitoring and control. Comput. Chem. Eng. 140.

Geladi, P., Kowalski, B.R., 1986. *Partial Least Squares Regression: a Tutorial*. Anal. Chim. Acta 185, 1–17.

Gurden, S.P., et al., 2001. A comparison of multiway regression and scaling methods. Chem. Intell. Lab. Syst. 59, 121–136.

Szymańska, E., et al., 2012. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics 8, S3–S16.

Indahl, U.G., Næs, T., 1998. Evaluation of alternative spectral feature extraction methods of textural images for multivariate modeling. J. Chemom. 12, 261–278.

Walczak, B., Massart, D.L., 2001. *Tutorial: dealing with missing data part I*. Chemom. Intell. Lab. Syst. 58, 15–27.

Suoza, F.A.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst. 152, 69–79.

Fortuna, L., Graziani, S., Xibilia, M.G., 2005. Soft sensors for product quality monitoring in debutanizer distillation columns. Control Eng. Pract. 13, 499–508.

Arteaga, F., Ferrer, A., 2002. Dealing with missing data in MSPC: several methods, different interpretations, some examples. J. Chemom. 16, 408–418.

Wang, D., Liu, J., Srinivasan, R., 2010. Data-driven soft sensor approach for quality prediction in a refining process. IEEE Trans. Ind. Inform. 6 (1), 11–17.

Varmuza, K., Filzmoser, P., 2009. Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, New York.

Eriksson, L., et al., 2013. Multi- and Megavariate Data Analysis Basic Principles and Applications, 1. Umetrics Academy.

Rajalahti, T., et al., 2009. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemom. Intell. Lab. Syst. 95 (1), 35–48.

Tran, T.N., et al., 2004. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). Chemom. Intell. Lab. Syst. 138 (15), 153–160.

Wang, Z.X., He, Q.P., Wang, J., 2015. Comparison of variable selection methods for PLS-based soft sensor modeling. J. Process. Control 26, 56–72.