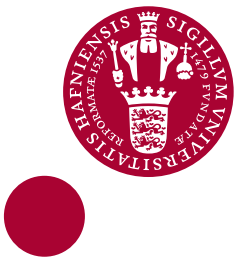


Modeling of the flux decline in a continuous ultrafiltration system with multiblock partial least squares

Klimkiewicz, Anna; Cervera-Padrell, Albert Emil; van der Berg, Franciscus Winfried J

Published in:
Industrial & Engineering Chemistry Research

DOI:
10.1021/acs.iecr.6b01241

Publication date:
2016

I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

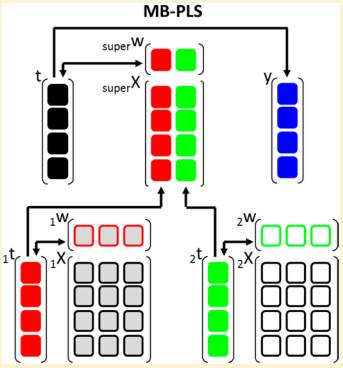
Modeling of the Flux Decline in a Continuous Ultrafiltration System with Multiblock Partial Least Squares

Anna Klimkiewicz,*[,†,‡] Albert Emil Cervera-Padrell,[†] and Frans van den Berg[‡]

[†]Novozymes A/S, 4400 Kalundborg, Denmark

[‡]Spectroscopy & Chemometrics section, Department of Food Science, Faculty of Science, University of Copenhagen, 1958 Frederiksberg C, Denmark

ABSTRACT: This study investigates flux decline in ultrafiltration as a capacity measure for the process. A continuous ultrafiltration is a multistage process where a considerable coupling between the stages is expected due to similar settings on the subsequent recirculation loops and recirculation of parts of the process streams. To explore the flux decline issue from an engineering perspective, two ways of organizing process signals into logical blocks are identified and used in a multiblock partial least-squares regression: (1) the "physical location" of the sensors on the process layout and (2) "engineering type of tags". Abnormal runs are removed iteratively from the original data set, and then the multiblock parameters are calculated based on the optimized regression model to determine the role of the different data building units in flux prediction. Both blocking alternatives are interpreted alongside offering a compact overview of the most important sections related to the flux decline. This way one can zoom in on the smaller sections of the process, which gives an optimization potential.



1. INTRODUCTION

Classical engineering strategies do not always perform well in control and optimization of full-scale biomanufacturing steps. This can be assigned to the complex multistage nature of these production systems, which cannot be described sufficiently accurately by mechanistic or first principle concepts.[1] The alternative, use of historical production records combined with statistical or data-driven models for process optimization, calls for apt empirical methods. Principal component analysis (PCA) and partial least squares (PLS) are popular multivariate dimension-reducing methods which are known to cope well with challenges associated with historical production databases, such as their enormous size, a high degree of correlation between variables, low signal-to-noise ratio, and recurrent missing values.[2]

In the case where the process measurements and signals originate from different phases in a manufacturing process, it is possible to improve the interpretability of multivariate models by multiblock methods.[2,3] They are an extension of well-known "single-block" factor models such as PCA and PLS. The popularity of multiblock methods has, however, grown only modestly over time. An explanation for this limited popularity is that orginally these methods were developed for improved (regression) modeling, but it was shown early on that most strategies are equivalent—in predictive performance—to PCA and PLS models on augmented data sets.[3] Instead, the important added "twist" of multiblock methods is the additional data organization layer plus block-specific information and diagnostics that they provide, which alleviates the risk of being overwhelmed by the size of the collected data set.[4] Some industrial applications for modeling and monitoring of production processes have been reported in the chemical,[5,6] pharmaceutical,[7,8] and food[4] sectors. The potential use of blocking is for the "same product" at different stages or phases of processing, such as distinctive time-steps in batch-wise production, as seen in, e.g., tablet production, or successive unit operations or sections of a unit in a continuous mode operation, as encountered in downstream bioprocessing.[1,9] The selection of a proper blocking structure for the process at hand is driven by the aim of the investigation and based on engineering intuition. Guidelines from the chemical process industry suggest that blocks should correspond as close as possible to discrete units of the process, in which all variables in one block are expected to be highly coupled, while there is less coupling expected between variables in neighboring unit operations.[6]

The multiblock PLS (MB-PLS) algorithm allows for the calculation of additional parameters such as so-called super level weights (the contribution of each data block to the solution), block level scores, and the percentage variation explained per data block. The advantage of the multiblock approach is that, by examining block contributions next to individual variable contributions, it eases the interpretation and helps in the understanding of the product and process analysis. The low-level block models can still be studied by their local block level scores and weights or loadings and the overall model (upper or super level) by the super level scores and

Received: April 1, 2016
Revised: August 30, 2016
Accepted: August 31, 2016
Published: September 1, 2016

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

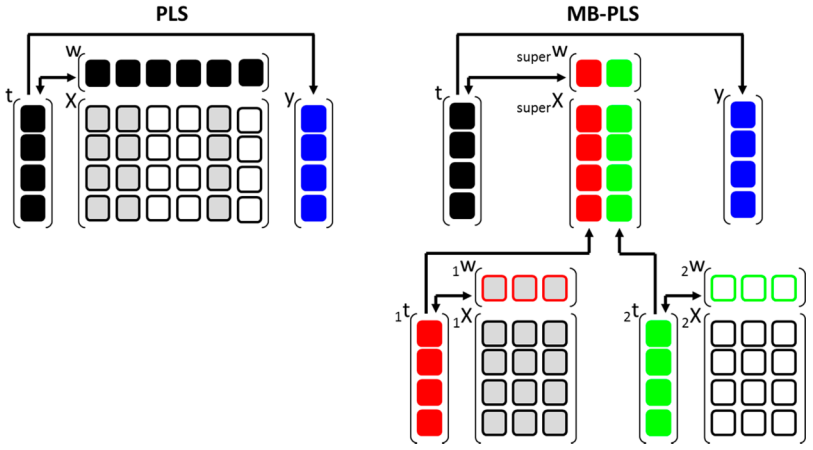


Figure 1. Conceptual scheme of the MB-PLS model.

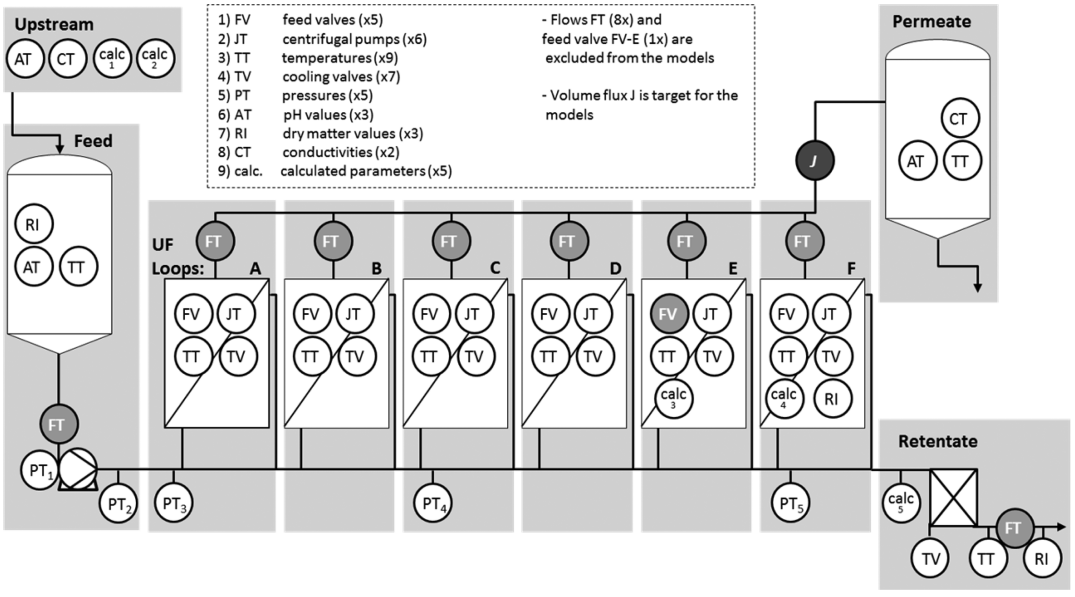


Figure 2. Scheme of the ultrafiltration system plus the approximate location of forty-nine process measurements and five calculated engineering values; flow signals and the throttling valve on loop E are excluded during modeling.

weights. Multiblock methods can thus be used to group process operating variables into meaningful blocks according to the operational phase and concern both the inner relationship within each phase and the interrelationship between different phases. This helps to identify the important parts of the process and, if necessary, to trace causes back to, e.g., the raw data.[3] Via the multiblock approach, one can build a model for the full process that will take into account the interactions between the units and their relative importance to the final product quality.[2]

There are three prevalent ways to obtain MB-PLS models.[10] The first method uses the block level scores for deflation of dependents **X** and independents **y**[11] in the regression equation $\mathbf{y} = \mathbf{X}\cdot\mathbf{b}$, which ensures orthogonality between the block level scores. In the second approach, the algorithm uses the super level scores to deflate **X** and **y**,[3,8] and it has proved to lead to a superior predictive performance. The results of the latter are equal to the calculation of the standard PLS on the combined or augmented matrix from all data blocks (Figure 1), providing the same weighing and variable scaling is applied.[3] This algorithm also works faster and proved to be better at handling missing values. In the third method, only **y** is deflated using the super level scores. This deflation scheme was recommended to prevent mixing up information at the block level, which in turn should lead to the easier interpretation of the block level scores.[10] For a detailed theoretical and algorithmic viewpoint, we recommend existing literature.[3,10,12] Data block scaling is an important issue in multiblock applications, comparable with variable scaling in regular bilinear modeling. Depending on the block scaling, quite different results, and hence interpretations, can be obtained.[13] Block weighing can be selected based, e.g., on the process knowledge or performance expectations. However, if no such information is available, all blocks should initially be given an equal contribution by scaling their variance to the equal sum-of-squares (so-called block normalization). This is especially important if the number of process variables in different blocks varies considerably. All in all, it can be a good strategy to try and investigate some different combinations of block weights and blocking in MB-PLS and compare the cross-validated prediction errors. If results are inferior to the standard PLS model with no block-weighing, then blocking is done incorrectly.[5,6]

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

This study uses flux in ultrafiltration (UF) as a capacity measure of a process and focuses on the block level to investigate the overall flux values, or more specific flux decline as a function of process time. Significant attention, in both public research and industry, has been paid to better understand the mechanisms of membrane fouling observed as flux decline in UF.[14−17] These problems clearly affect the production scheduling and hence economics in downstream biomanufacturing. In the Novozymes production facilities at Kalundborg (Denmark), a project was initiated to investigate the flux decline issue based on historic full-scale processing data. Preceding exploratory studies directed our attention to one of the manufacturing recipes which is characterized by a very steep flux decline.[14] In the current study, we look closer at this specific group of production runs and treat it as a regression problem. Specifically, we want to construct models based on process data to predict the flux values and we want to interpret the role of the different data building blocks in this prediction.

2. MATERIALS AND METHODS

2.1. The UF System. A plate and frame ultrafiltration system is operated as a multistage recirculation plant where the smallest working element of the UF equipment is a membrane (Figure 2).[18] Membranes retain enzyme molecules (based on their size and shape) in the retentate while allowing for the permeation of water and small molecules. Membranes are polymer sheets, fitted in pairs between supporting hard plastic plates with spacer channels. The pores of the ultrafiltration membrane are very small, and a pressure must thus be applied to make the separation process effective. The feed is pumped between the paired membranes flowing parallel to the membrane surface while permeate has a transverse flow direction (termed "cross-flow"). This type of process flow minimizes fouling and excessive material build-up. The permeate passes through the membranes into the plastic plates spacers, where it is led away through a permeate tube. One membrane module consists of hundreds of membrane sheets and supporting structures. Several modules working in parallel form a recirculation loop. These stages are called "loops" in Figure 2.

Each recirculation loop is supplied by a centrifugal pump (JT) and the accompanying throttling valve (FV) to provide pressure and to ensure an adequate cross-flow velocity of the feed over the membranes. This helps permeate to pass through the membranes, provides a fresh flow of the feed and recirculation liquid, and prevents too much concentration polarization over the membrane area. Centrifugal pumps generate heat which has to be removed by cooling (TV). Other key components external to the loops are a feed tank followed by the feed pump (PT_1), a permeate tank, pipelines, and a heat exchanger on the retentate stream. There is also a number of flow transmitters (FT) installed to monitor and control the throughput.

A membrane system designed as a multistage recirculation plant with a high volumetric concentration ratio must be controlled based on a very small flow of the retentate.[18] There are two main control modes available. The first one is using the concentration of dissolved solids measured by a refractometer (RI) located on the last recirculation loop. As soon as the concentration is equal to or exceeds an RI value set by the operator, the regulation valve opens and adjusts its position during filtration to ensure the desired enzyme concentration in the retentate stream. As a second option, concentration can be controlled using a flow ratio between the volume entering the plant and the volume of retentate leaving the plant. This calculated parameter is called the volumetric concentration degree, and it is labeled as "calc5" in Figure 2. Additional "upstream" information, related to the primary separation of the enzyme from the biomass, is used in this study. It covers parameters such as pH (AT), conductivity (CT), dilution (calc1), and dosing of the flocculation chemical (calc2).

It is not easy to track the path of a product/effluent stream in this UF operation. In general, recirculation loops work in sequence from A to F but the retention times on each loop or even within the entire unit are not known. The proper lags between different process signals would as a consequence be extremely hard to determine because they vary owing to the different number of the loops in use, the degree of recirculation on the loops, the process temperatures, the properties of the feed, the degree of membrane fouling, and the degree of up-concentration. Moreover, process signals have different logging frequencies on the data historian, and it is not expected that shifting the signals to match with a minute precision would make any significant difference. Instead, we use average values over a fixed and equidistant time interval and no lagging for any of the parameters. Additionally, also the reference value in this study, the volume flux, is a weighed estimate based on the permeate flow over the same time interval.

The UF system can only run for a limited period before the membranes have to be cleaned. In daily practice, the UF capacity is monitored based on the permeate flow out of the UF loops and the retentate flow (FT's in Figure 2). The operator stops the unit and proceeds to cleaning when these parameters drop to unacceptably low values. It is, however, problematic to use these seven parameters for the postrun capacity evaluation, especially since not all UF loops are in use all the time. Instead, we calculate the volume flux (J, $L{\cdot}m^{-2}{\cdot}h^{-1}$) by relating the total permeate flow to the working membrane area at every timestamp. This is done according to the formula:

$$J(t) = \frac{v_{tot}(t)\cdot 1000}{Wb(t)\cdot A} \tag{1}$$

where $v_{tot}(t)$ is total permeate flow, summed values from all loops, at time t, in $m^3{\cdot}h^{-1}$, 1000 is the adjustment for L instead of m^3, $Wb(t)$ is the number of loops working at timestamp t, based on the assumption that a loop is working if the power of the corresponding centrifugal pump is larger than 1%, and A is the membrane area corresponding to one loop (m^2). It should be noted here that all values, including time, have been scaled to arbitrary units to mask proprietary information.

2.2. Structure of the Data Set. All data originates from an ultrafiltration operation in a full-scale downstream process of industrial enzymes. The process data set is a sample from records registered over a year of production of one type of intermediate enzyme product. A previous study[14] brought our attention to the processing variant which was associated with a particularly rapid membrane fouling (called "recipe 3" in ref 14). Consequently, this group of production runs ($I = 40$) is in the center of the follow-up investigation presented here. As in the previous study, it was decided to analyze only the data corresponding to the (quasi-)steady-state UF phase after exclusion of the startup phase. The term "process tag" is used throughout this study as a synonym for process signal or variable; in the production environment it is used in reference

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

to process operating variables which are sampled and stored in the data historian. Forty-nine tags are physically installed near the locations depicted in the UF diagram in Figure 2. Eight tags from flow meters are excluded from the analysis, as they are either used in the calculation of flux or conjugated to it, owing to the regulation of concentration factor and pressure in the unit. The "FV" tag of loop E is also excluded, as its value does not vary across the data set. In addition, five meaningful engineering parameters have been calculated based on the tags shown in Figure 2 and some other process variables not revealed. Hence, each UF run i (i = 1, ..., I) is represented by a data matrix $\mathbf{X}_i$ with N_i measurement occasions (timestamps) by J variables (J = 45). J's are average values over a fixed and equidistant time interval of the original process operating variables. The total number of timestamps over all data sets is equal to $N = \sum_{i=1}^{I} N_i$ = 623. Process measurements recorded upstream have been compressed or expanded using linear interpolation in the time dimension to match the length of the corresponding UF run.

Flux (J, in Figure 2) is used as the dependent $\mathbf{y}$-variable. Figure 3 illustrates the variation in the flux profiles encountered

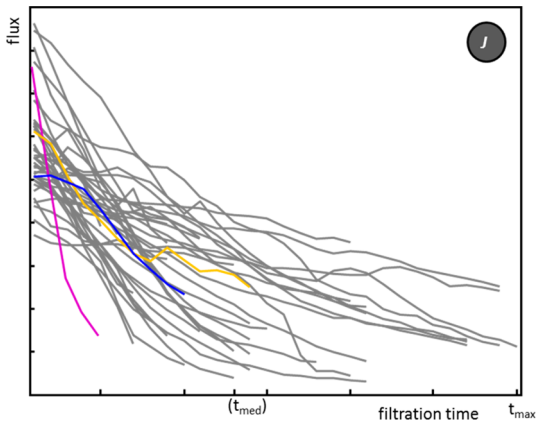


Figure 3. Flux values/fouling profiles encountered in the investigated data set (I = 40).

in the examined data set. These flux reduction profiles might at first encounter resemble trajectories typically seen in batch processes. Nevertheless, the perfect development in a steady-state continuous UF process is expected to be a plateau, preferably situated at a high flux level. One can also recognize that runs significantly vary in length as filtration is stopped either due to unacceptably low flux or because the order (a "lot of material") has been processed.[14]

2.3. Data Analysis. We identify and compare two concepts for organizing process signals into logical blocks. The first blocking strategy is to group variables according to the "physical location" of the sensors with respect to the process layout as indicated by the shaded areas in Figure 2. This resulted in the formation of ten blocks: (1) upstream parameters, (2) feed, (3−8) recirculation loops A−F,(9) retentate, and (10) permeate. In this scenario, there are between three and seven process variables per block. In the second approach variables are grouped according to the "engineering type of tags" (clustering together variables of similar characteristics, e.g. readings from temperature sensors, records from the centrifugal pumps, etc.). This blocking strategy leads to the formation of nine groups as listed in the frame presented in Figure 2, comprising between two and nine variables.

Multiblock PLS with super level scores deflation of $\mathbf{X}$ and $\mathbf{y}$ has been used throughout this work, the general structure of which is depicted in Figure 1.[10] In our computations first the standard PLS models are calculated and examined. Next, the multiblock parameters are determined from the optimized PLS model interpretation using the super level scores to deflate $\mathbf{X}$ and $\mathbf{y}$.[3] Strategy with only $\mathbf{y}$-deflation[10,19] was also investigated. The outcome was very similar as when both $\mathbf{X}$- and $\mathbf{y}$ were deflated, leading to exactly the same interpretation, and hence, it is not presented here. Data analysis was performed using Matlab (version 8.0.0.783/R2014a, Mathworks, USA) in combination with in-house code and the PLS Toolbox (Version 7.9.5, Eigenvector Research Inc., Manson, WA, USA).

3. RESULTS AND DISCUSSION

3.1. PLS Model on Augmented Data. Steady-state filtration data from the 40 runs has been concatenated in the process tag direction (variable-wise), and all data has been autoscaled. A PLS model is built between the process variables $\mathbf{X}$ (N×J) and the flux $\mathbf{y}$ (N×1). This way, every sampling time is represented by a row vector of length equal to the number of process variables and the number of rows is determined by the time-horizon included in modeling (N). PLS models extract latent variables (LVs) that explain the variation in the process data $\mathbf{X}$ which is most predictive of flux and disregard the measurement errors and random variations which are uncorrelated with other $\mathbf{X}$-variables and the flux. Stratified cross-validation has been applied to determine the optimal number of latent factors in the model where each of the investigated runs (I) is assigned a number between 1 and 10 and four runs with the same numbers were removed at the time. The average root mean squared error of cross-validation (RMSECV) as a function of model complexity is plotted in Figure 4 together with the root mean squared error of

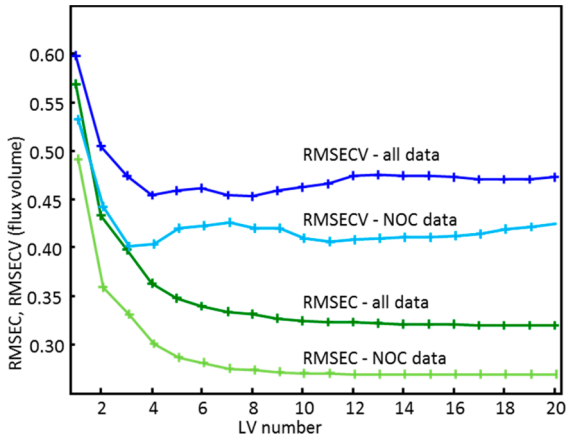


Figure 4. Calibration (RMSEC) and cross-validation (RMSECV) errors for the PLS models constructed using all data (I = 40, N = 623) or NOC data (I = 30, N = 508).

calibration (RMSEC). Results of cross-validation suggest that the model works best using four LVs which corresponds to explaining 86.8% flux variation and an RMSECV equal to 0.46 (arbitrary units). The first dimensions of the PLS model are certainly the most dominant but for prediction purposes all dimensions determined via cross-validation should be used.[6]

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

Next, four data points (meaning four time stamps) which appear extreme in the influence plot (not shown) have been removed. Samples showing the outstanding behavior were either from the very beginning or the end of a process run. In these instances, pressure, which is normally tightly controlled during UF, had been outside its normal limits. Removal of the outlying data points did not affect the decision on the number of LVs in the model which now corresponds to an RMSECV equal to 0.45. The first LV explains the highest amount of variation in **y** (68.8%) and the subsequent components explain significantly less variation of the flux (13.1%, 3.4%, and 1.9%, respectively).

Projection of the process time-points on the first two latent factors reveals which runs follow the normal operating conditions (NOC). Based on the LV1 vs LV2 score plot, it is possible to classify the behavior of the process as NOC or AOC (abnormal operating conditions). In Figure 5, the behaviors of

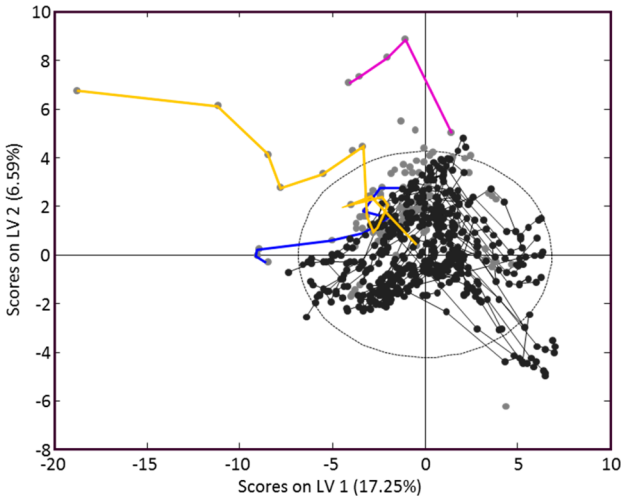


Figure 5. LV1 vs LV2 score plot of the PLS model constructed on NOC data; three instances of AOC runs are projected onto the model: blue and yellow represent abnormal pressure behavior, purple shows abnormal temperature behavior (see text for details).

what were iteratively identified as the normal processes are marked in gray on the LV1 vs LV2 score plot. Runs classified as NOC start on the right side of the plot (high positive score) and end on the left side of the plot (high negative score), and they follow an arc-shaped trajectory. Thus, the first LV represents mainly filtration time. The second LV (dictated by the "stretch" of an arc) is related to the regulation on the first two UF loops (A and B in Figure 2). Those data points which are located outside the 95% coverage ellipse and which are characterized by a high score on LV1 and a very low score on LV2 represent the situation when not all loops are used at the early stages of a UF run. This explanation could be found by constructing the Hotelling's T^2 contributions plot for those timestamps (not shown) and confirmed by looking at the raw data (not shown). Specifically, the Hotelling's T^2 contributions from the process sensors located on the first two loops were high. It is not unusual to run with a lower capacity at the early stages of ultrafiltration and add loops gradually over the course of a process run. Furthermore, recirculation loops which are physically located as first are added as the last. Therefore, these data points which are characterized by a high score on the first LV and a very low score on the second LV have been kept in the model. Ten AOC runs have been removed iteratively from the original data set, as they follow a distinctly different trajectory to those of the NOC runs. It should be emphasized here that all runs investigated are within predetermined quality control limits. All the "abnormal runs" are associated with optimization trials, operator interventions, or other known causes. However, since our aim in this study is to elucidate the relationship between process variables and regular permeate flux decline it was decided to model NOC runs only. A new PLS model was calculated using the remaining 30 runs. The stratified cross-validation procedure points at three LVs as the optimal number of components in this model (Figure 4) which corresponds to an RMSECV equal to 0.40. Components used explained jointly 89.0% of the variation in **y** and 35.9% of the variation in **X**.

For a diagnostic interpretation Figure 5 includes three instances of AOC runs which were projected onto the model built using NOC data only. In general, the reason for the outstanding behavior of the excluded runs could be related to either a noticeable drift in the ultrafiltration pressure or extreme temperature values. Abnormal events manifested themselves along all three latent components. If the fault was related to pressure, then it showed itself across LV1. If the unusual behavior was caused by extreme temperatures, then it could be identified across LV2. Two examples of the first situation are seen in Figure 5. This could be confirmed in the raw signals involved in pressure regulation and monitoring as plotted in Figure 6. Pressure is the driving force in a membrane filtration system.[18] Those measurements and the regulating pump are strongly correlated owing to a stringent control over the ultrafiltration pressure. Variable "PT_4" (Figure 6) is the tag directly controlled using the feed pump ("PT_1", Figure 6). The three remaining pressure sensors are only monitored and not

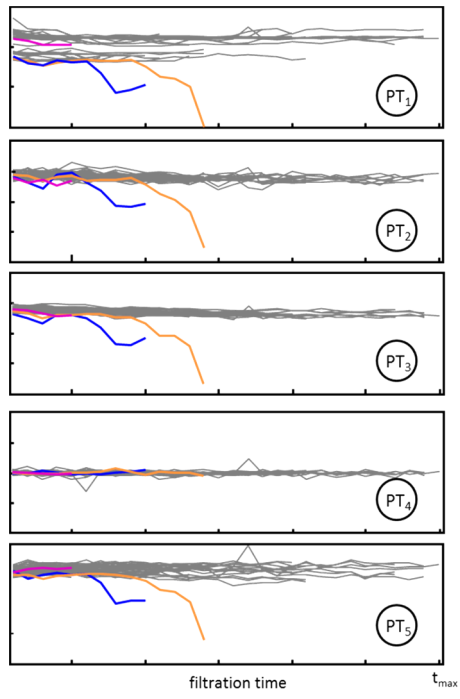


Figure 6. Signals related to pressure control (see Figure 2) in the UF system collected during NOC runs (marked in gray) and examples of the AOC runs: pressure related (blue and yellow) and temperature related (purple; compare with Figure 5).

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

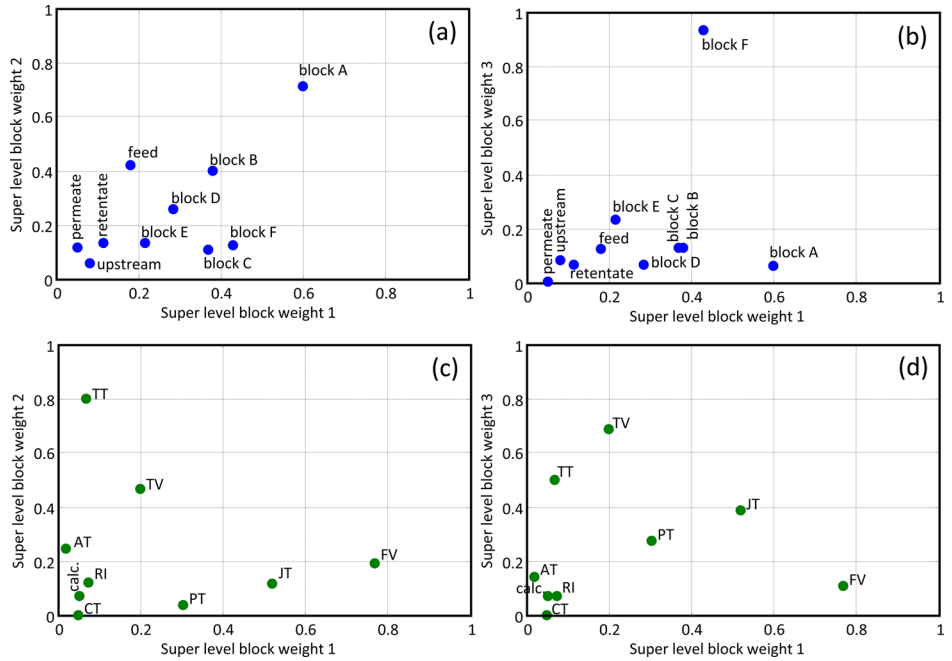


Figure 7. MB-PLS model super level block weights when blocking is done according to (a, b) physical location or (c, d) sensor type.

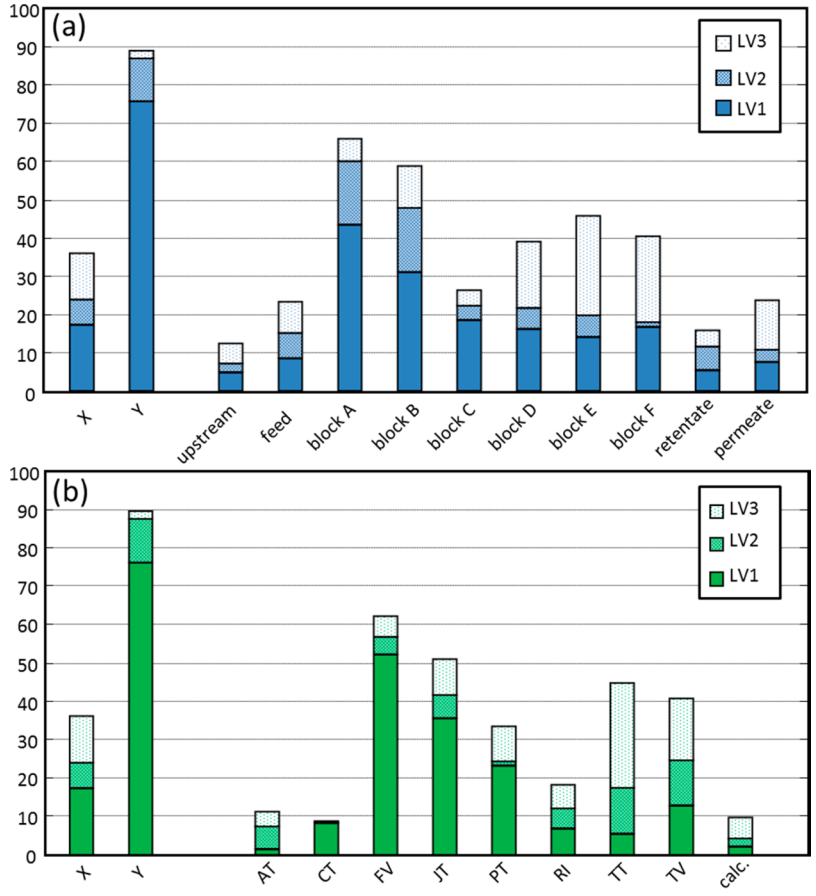


Figure 8. MB-PLS model variances explained when blocking is considered according to (a) physical location or (b) sensor type (see Figure 2 for interpretation).

used in the closed loop feedback control. From this overview of part of the raw data, it is clear that pressure at "PT_4" is always within its control limits. However, a clear decline in pressure at other measuring points happened over the filtration time in the case of the above-mentioned runs. It is an interesting observation that even though pressure at "PT_4" is always tightly and effectively controlled, pressure at the other measuring points shows a strong decline in those runs. Also

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

marked in Figure 5 (and in Figure 6 for completeness) is an example of a situation when the processing temperature was controlled significantly higher than usual.

It is interesting to note that in the case of the examined process the first two LVs would be truly sufficient for the monitoring purposes. This observation is consistent with the recommendation made before for a large chemical process.[6]

3.2. Blocking in MB-PLS. Two MB-PLS models have been calculated from the optimized PLS model (I = 30 runs, 3 LVs). They differ in the way that variables are arranged in conceptually meaningful blocks using system knowledge and engineering insight (Figure 2). The objective is to keep track of different blocks during the analysis which leads to a more parsimonious investigation compared to keeping track of individual variables. Block normalization was also investigated. However, it was concluded that for our process it might force the solution in a direction where the fact-finding aspect of the MB-PLS models is suppressed. Moreover, no significant improvement in terms of RMSECV was registered when blocks entered the model with equal norm. This is a natural consequence of the different building blocks being not too different in size (ranging from two to eight variables). It should be noted that if block normalization would have been used the relation to the full model would be lost and the two blocking approaches would not be comparable. This would certainly make the model interpretation more complex. Therefore, the two MB-PLS models presented in this study both have as starting point the optimized PLS model described in the previous section with each process tag having a weight one (due to the autoscaling preprocessing).

Figure 7 presents the super level block weights of the two MB-PLS model calculated when sectioning is done according to physical location or to the sensor type. Figure 8 summarizes the variances explained per block for each LV retained in the model, again for the two blocking strategies. It is expected that weights and variances explained point at similar phenomena on the corresponding LV. From a phenomenological point of view weights represent features in the process data $\mathbf{X}$ which are related to the original flux values in $\mathbf{y}$. In both blocking strategies a solid coupling between variables in each block is anticipated. In the case of physical blocking also a considerable coupling between the sections is expected. On the other hand, each section can face its own set of distinctive events like membrane fouling (or e.g. more extreme upsets like leakage). It is, therefore, rational to split up the process into physical blocks and keep track of these sections separately. We found it most useful when both blocking alternatives are interpreted together rather than choosing one over the other. For instance, super level block weights on the first LV point at loops A, F, B, and C and at the feed valves and centrifugal pumps. To translate this into process understanding, cross-flow regulation on the A, B, C, and F loops can explain 75.7% of the variation in flux. The second LV explains 11.3% of the variation in $\mathbf{y}$, and implying from the super level weights this can be related to the temperature regulation on loops A, B and in the feed.

The third LV is explaining only 2.0% of the variation in flux. Judging from the variances explained by this component, it is primarily related to temperature regulation on loops D, E, and F (Figure 8a). Yet, super level weights point at loop F being the most important (Figure 7b).

Proper expert insight is necessary to clarify which of the observed relations lead to new, unexpected findings and hence can be used in optimization. After a closer inspection, a dominant amount of the $\mathbf{y}$ variation covered by the first two LVs can be explained by the mechanics of the UF system. For instance, the cross-flow control on the recirculation loops is indirectly responding to the degree of membrane fouling, hence, to the flux decline. Most of the process signals dominant on the first two LVs cannot be utilized to improve process performance. In relation to the high variance explained by the first LV, it is, however, interesting to have a closer look at the "PT" tags which are third in terms of super level weight on this component. Feed pressure (PT_3) to the unit shows a decrease over filtration run time, and it correlates positively to flux decline (Figure 9, R^2 = 0.67, in relation to flux data shown in Figure 3). This observation would be very hard to make just by looking at the raw data before the AOC runs have been removed (Figure 6). Interestingly, except for variable "PT_4" which is the tag used in control, the other pressure monitoring points show a drift over the filtration time. It could be an indication that the current control strategy is not optimal and that controlling the pressure using the other measurements in a cascade setting may result in a more stable overall flux. This was not revealed in our previous data mining approaches of a more diversified data set where pressure tags quickly fall out of analysis as they appeared constant over the course of ultrafiltration.[14]

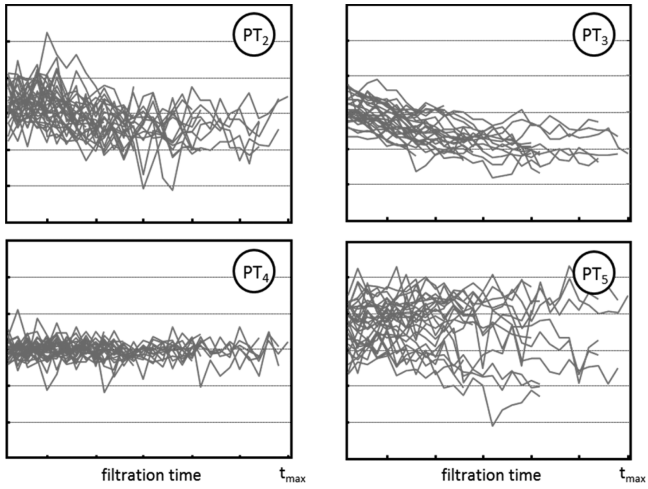


Figure 9. Pressure measured during the NOC runs (see Figure 2).

A second interesting observation from the MB-PLS models is related to the third LV which, as was noted before, is associated with temperature regulation in the last three recirculation loops. Originally, temperatures have been at the same level on all loops (represented by the red markers in Figure 10). At some point in time it was decided to lower the temperature of the last three loops (represented by green markers in Figure 10), and it can be observed that most of the AOC runs belonged to the first group (shaded markers in Figure 10). The two processing recipe variants overlap on the first two LVs (Figure 10a). On the other hand, Figure 10b shows that the third LV neatly separates data according to the temperature on the last three recirculation loops. This distinction is naturally even more striking on temperature block level scores on the third LV (not shown). Indisputably, the cross-validated PLS model (Figure 4) points at a third LV as being important for the flux prediction. Understanding this relation is however not straightforward from the PLS scores and loading plots but can be explained based on the more parsimonious representation of the MB-PLS

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

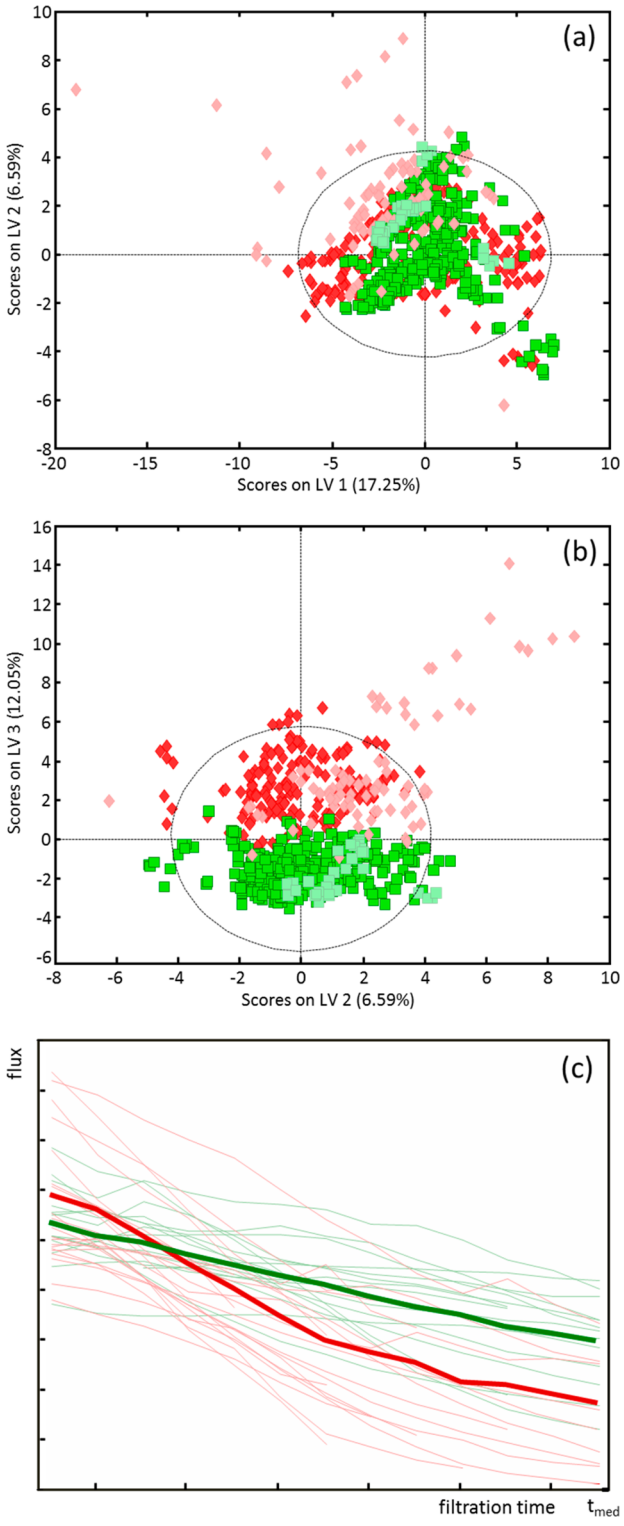


Figure 10. (a) LV1 vs LV2 (analog to Figure 5) and (b) LV1 vs LV3 score plots of the PLS model constructed on NOC data colored according to a temperature related processing recipe change: red markers, equal temperature on all loops; green markers, lower temperature on D-F loops; shaded markers, AOC (excluded) data projected into the NOC PLS model; (c) mean (bold) and individual (thin) flux reduction profiles encountered until median ultrafiltration length, colored according to the recipe.

models and, afterward, on the low-level interpretation of the raw data. If we compare the mean flux trajectories up to the median filtration length of the investigated data set, it can be seen that the higher processing temperature converts into higher flux at the start of a run but a lower flux toward the end (Figure 10c). This is also reflected in the first LV as the score values of the runs with equal temperatures on all blocks are more spread on this latent component (Figure 10a). This corresponds to faster membrane fouling at the higher temperatures. It is important to recall here that optimization of a UF system is combined mechanistic and stochastic challenge. The MB-PLS model identified more mechanistic (run time along LV1 in Figure 10a) or operational principles (temperature recipe along LV3 in Figure 10b). But Figure 10c shows that for the same settings the flux decline profiles still differ significantly. Hence, next to operational recipe and local closed-loop control strategies there are obvious opportunities to improve the performance of a complex system like ultrafiltration by data-driven statistical process control.

4. CONCLUSIONS

The interpretability of the PLS model can be more holistic and simplified by calculation of the lower and super level multiblock parameters. In the approach taken by us, MB-PLS is not a different variant of the PLS model, but an additional set of diagnostics offering a prompt overview of the most important phenomena happening in the data. In the investigated process, we identify two natural ways to block the data and find it most useful to use them together. As process variables were assigned to groups corresponding to distinct phases of the process or belonging to similar engineering types of sensors, it was considerably easier to study and interpret the behavior of these blocks rather than keeping track of forty-five individual loading values. The upper level of the MB-PLS indicates the relationship between different groups of variables and points at those which are the most relevant in the prediction of flux and flux decline. This multiblock feature is helpful in concentrating efforts of the process engineers on those areas that have an optimization potential.

Similarly to our previous study, it appears that higher processing temperature can have both positive and negative consequences to the UF flux. Additionally, a potential field for improvement has been reported, related to the pressure monitoring point used in the closed loop feedback control of ultrafiltration pressure.

In a previous manuscript,[14] we have used blocking in the row or time direction to look at differences and similarities between and within process runs. It could be a useful future perspective to develop methods capable of blocking in both the row and column directions—hence, time or dynamics and equipment layout—which in turn could relax the exploration and analysis of the multivariate historical data sets even more.

AUTHOR INFORMATION

Corresponding Author

*E-mail address: anna.k@food.ku.dk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to acknowledge an Industrial Ph.D. grant from Innovation Fund Denmark. Process engineers, scientists, and technicians at the Novozymes recovery plant and downstream optimization department in Kalundborg, Den-

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698

mark, are acknowledged for providing invaluable advice and suggestions along the way.

REFERENCES

(1) Camacho, J.; Picó, J.; Ferrer, A. Multi-phase analysis framework for handling batch process data. *J. Chemom.* **2008**, *22*, 632.
(2) Kourti, T. Process analysis and abnormal situation detection: from theory to practice. *Control Systems, IEEE* **2002***22*, 10.10.1109/MCS.2002.1035214
(3) Westerhuis, J. A.; Kourti, T.; MacGregor, J. F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **1998**, *12*, 301.
(4) Bro, R.; van den Berg, F. W. J.; Thybo, A.; Andersen, C. M.; Jørgensen, B. M.; Andersen, H. Multivariate data analysis as a tool in advanced quality monitoring in the food production chain. *Trends Food Sci. Technol.* **2002**, *13*, 235−244.
(5) Kourti, T.; Nomikos, P.; MacGregor, J. F. Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS. *J. Process Control* **1995**, *5*, 277.
(6) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process monitoring and diagnosis by multiblock PLS methods. *AIChE J.* **1994**, *40*, 826.
(7) Lopes, J. A.; Menezes, J. C.; Westerhuis, J. A.; Smilde, A. K. Multiblock PLS analysis of an industrial pharmaceutical process. *Biotechnol. Bioeng.* **2002**, *80*, 419.
(8) Westerhuis, J. A.; Coenegracht, P. M. Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *J. Chemom.* **1997**, *11*, 379.
(9) Yao, Y.; Gao, F. Phase and transition based batch process modeling and online monitoring. *J. Process Control* **2009**, *19*, 816.
(10) Westerhuis, J. A.; Smilde, A. K. Deflation in multiblock PLS. *J. Chemom.* **2001**, *15*, 485.
(11) Wangen, L.; Kowalski, B. A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemom.* **1989**, *3*, 3.
(12) Qin, S. J.; Valle, S.; Piovoso, M. J. On unifying multiblock analysis with application to decentralized process monitoring. *J. Chemom.* **2001**, *15*, 715.
(13) Kreutzmann, S.; Svensson, V. T.; Thybo, A. K.; Bro, R.; Petersen, M. A. Prediction of sensory quality in raw carrots (Daucus carota L.) using multi-block LS-ParPLS. *Food Quality and Preference* **2008**, *19*, 609.
(14) Klimkiewicz, A.; Cervera-Padrell, A. E.; van den Berg, F. W. J. Multilevel Modeling for Data Mining of Downstream Bio-Industrial Processes. *Chemom. Intell. Lab. Syst.* **2016**, *154*, 62.
(15) van den Berg, G.; Smolders, C. Flux decline in ultrafiltration processes. *Desalination* **1990**, *77*, 101.
(16) Marshall, A.; Munro, P.; Trägårdh, G. The effect of protein fouling in microfiltration and ultrafiltration on permeate flux, protein retention and selectivity: a literature review. *Desalination* **1993**, *91*, 65.
(17) Linkhorst, J.; Lewis, W. J. Workshop on membrane fouling and monitoring: a summary. *Desalin. Water Treat.* **2013**, *51*, 6401.
(18) Wagner, J. *Membrane Filtration Handbook: Practical Tips and Hints*; Osmonics: Minnetonka, MN, 2001.
(19) Dayal, B.; MacGregor, J. F. Improved PLS algorithms. *J. Chemom.* **1997**, *11.1*, 73.

DOI: 10.1021/acs.iecr.6b01241
Ind. Eng. Chem. Res. 2016, 55, 10690−10698