UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Multivariate Statistical Process Optimization in the Industrial Production of Enzymes

Industrial Ph.D. thesis by
Anna Klimkiewicz
• 2016 •

novozymes®
Rethink Tomorrow

UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Multivariate Statistical Process Optimization in the Industrial Production of Enzymes

Industrial Ph.D. thesis by
Anna Klimkiewicz
▪ 2016 ▪

Title

Multivariate Statistical Process Optimization in the Industrial Production of Enzymes

Author

Anna Klimkiewicz

Date of Submission

29 February 2016

Date of Defense

21 April 2016

Supervisors

Assoc. Prof. Frans W.J. van den Berg (*principal supervisor*)
Spectroscopy & Chemometrics
Department of Food Science, Faculty of Science
University of Copenhagen

Research Scientist Albert E. Cervera-Padrell, Ph.D. (*company supervisor*)
Solid Products Development
Novozymes A/S, Denmark

Assessment Committee

Assoc. Prof. Thomas Skov (*chairman*)
Spectroscopy & Chemometrics
Department of Food Science, Faculty of Science
University of Copenhagen

Principal Research Scientist Onno de Noord,
Statistics & Chemometrics,
Shell Global Solutions International BV, The Netherlands

Prof. Krist Gernaey,
Department of Chemical and Biochemical Engineering
Technical University of Denmark (DTU)

Cover illustration
Beta Renewables, cellulosic ethanol facility in Crescentino, Italy

ISBN 978-87-7611-994-2

Printed by SL grafik, Frederiksberg C, Denmark (www.slgrafik.dk)

Preface

This dissertation is submitted as a requirement for obtaining the Ph.D. degree at the University of Copenhagen. The presented work has been carried out at Multi-Purpose Production (MPP) and Downstream Optimization Departments, Novozymes A/S, and at the Spectroscopy & Chemometrics group (SPECC), Department of Food Science, Faculty of Science, University of Copenhagen. The study has been supervised mainly by Associated Professor Frans van den Berg, and has been funded by Innovation Fund Denmark and Novozymes A/S.

I am utterly grateful to my supervisor, Frans, for sharing his extensive knowledge, experience and ideas with me. He has been always available to help and able to push things in the right direction. I am very grateful to my co-supervisors Peter Mortensen and Christian Zachariassen for giving me the opportunity to perform this industrial Ph.D., for challenging me along the way but also for believing in me and remaining good friends. I am also very grateful to my other co-supervisors and managers: Marianne Becker Rousing for lessons in communication, Albert Cervera-Padrell and Mads Thaysen for inspiring and critical discussions on multivariate statistics and their help in making this research available to a wider audience. I am also indebted to Niels Murmann for initiation of this project as well as Steen Skærbæk and Morten Carlsen for further supporting this project.

As a part of the research, I had a three-month stay at Multivariate Statistical Engineering Group, Department of Applied Statistics, at the Technical University of Valencia. I am indebted to my supervisor during the stay Professor Alberto Ferrer for his kind help and introducing me to the area of batch process modeling. Furthermore, Raffa, Dani, José Manuel, Alberto, Abel and Marina are also thanked for making my stay a good experience both socially and scientifically.

I am thankful to all my colleagues at Novozymes and SPECC for contributing to a pleasant, welcome and professionally fruitful working environment. At Novozymes credits go to Boris, Ole, Jeff, Poul, Mark, Morten, Lars, Nis, Mohammad, Nikolaj, Sandra, Simon and Carsten. The combination of your various professional competences and personal qualities has provided me with the vivid and inspiring working environment. At SPECC special thanks go to Tine and Carina, for many good chats and helping with Danish version of the abstract, Parvaneh and Jannie for being such pleasant officemates, and Åsmund, for keeping me informed about interesting work-related opportunities including this Ph.D. project and for being a good friend. To all SPECC – thank you for making daily life, conferences and travelling so much fun!

A lot of love to my family and friends. Above all, I would like to thank my friend Agnieszka for her support, friendship and 'zen'-approach. And finally, I would like to thank my partner Marek for his patience and understanding. Thanks for keeping up with my work-centered attitude and for being there for me.

Anna Klimkiewicz

February 2016

Summary

In modern biotech production, a massive number of diverse measurements, with a broad diversity in information content and quality, is stored in data historians. The potential of this enormous amount of data is currently under-employed in process optimization efforts. This is a result of the demanding steps required in thoughtful data retrieval from the historian and the subsequent data pre-processing steps. Furthermore, efficient methods are needed capable of handling the data in the natural structure in which it was generated.

This dissertation work is meant to address some of the challenges and difficulties related to 'recycling' of historical data from full-scale manufacturing of industrial enzymes. First, the crucial and tedious step of retrieving the data from the systems is presented. The prerequisites that need to be comprehended are discussed, such as sensors accuracy and reliability, aspects related to the actual measuring frequency and non-equidistance retaining strategies in data storage. Different regimes of data extraction can be employed, and some might introduce undesirable artifacts in the final analysis results (POSTER II[1]). Several signal processing techniques are also briefly discussed and examples of applications presented, e.g. how to compensate for sensors with low signal to noise ratio or the handling of artifacts in the data. A second important step is alignment and synchronization of process data. This is particularly significant when looking at the relation between sequences of unit operations separated in time and, even more so when working with (semi-) continuous processes when generating the time series data. For this application, the potential of auto- and cross-correlation analysis and the effect of the prerequisite signal de-trending are explored in the context of the continuous granulation-drying process (POSTER I).

[1] Posters and papers marked by all-capitals can be found at the end of this thesis

The research presented in this thesis is primarily centered on the ultrafiltration step during which enzymes are purified and up-concentrated. The throughput of a continuous ultrafiltration operation is limited by the membrane fouling phenomena where the production capacity - monitored as flow through the membrane or flux - decreases over time. The flux varies considerably from run to run within the same product and likewise between different products. This variability clearly affects the production scheduling and leads to additional costs due to the more frequent membrane cleaning. The dataset examined in this investigation was compiled from records of conventional, univariate process sensors collected over several years of production of one type of intermediate enzyme products. Different strategies for the organization of these datasets, with varying number of timestamps, into data structures fit for latent variable (LV) modeling, have been compared. The ultimate aim of the data mining steps is the construction of statistical 'soft models' which capture the principle or latent behavior of the system under investigation. If this leads to new knowledge, it could be used for optimization of future production runs. Data reduced to mean value per run, combined with some other relevant features, has been used together with PLS2 regression in the primary investigation. It allowed us to identify the major differences between the processing variants of the investigated enzyme. Data arrangement into three-way cubes has been achieved by limiting the datasets to the median length. Studies with LV techniques after the batch-wise unfolding did not led to any special findings. Hence, it has been concluded that the process can be modeled sufficiently well when the datasets are concatenated variable-wise. The later studies used this type of data arrangement and focused only on the products with higher concentration degree as in those cases the flux decline problem has been the most pronounced. Blocking in the row or time direction was used in PAPER II. The dataset has a natural multilevel structure with level one being the process timestamps which are nested within the ultrafiltration runs, referred to as level two. Multilevel Simultaneous Component Analysis with invariant Pattern (MSCA-P) is applied to explore this historical dataset in the context of flux decline. We build on the two-level idea and expand the model to a third level: 'processing recipe'. In PAPER III blocking in the column or process tags direction has been used. A multiblock PLS breaks the process variables into smaller groups, clustering variables of similar importance and characteristics, to facilitate the

diagnostic procedure. Both methods lead to decomposition of the data structures into intuitively interpretable solutions by keeping the natural structure of the analyzed data.

Additionally, the ultrafiltration system has been also investigated in terms of product yield. The potential of NIR technology to monitor the activity of the enzyme has been the subject of a feasibility study presented in PAPER I. It included (a) evaluation on which of the two real-time NIR flow cell configurations is the preferred arrangement for monitoring of the retentate stream downstream to the UF, and (b) if the system can be used for statistical process monitoring and early warning/fault detection. It was possible to develop satisfying robust calibration models for four types of enzyme products where specific enzyme activities have been standardized into one global QC parameter. Finally, the study revealed that the less demanding in-line flow cell setup outperformed the on-line arrangement. The former worked satisfactory robust towards different products (amylases and proteases) and associated processing parameters such temperature and processing speed.

This dissertation work shows that chemometric methods specially designed for two-way and multiset problems have great potential as PAT tools as they fulfill the primary goal of PAT, namely to obtain a better process understanding in a faster and more intuitive way, especially when preserving the original data structure and dimensionality.

Resumé

I moderne bioteknologisk produktion er et massivt antal af forskellige målinger, med stor variation i informationsindhold og kvalitet, gemt i historikdata. Potentialet i denne enorme mængde data er i øjeblikket ikke udnyttet til fulde i procesoptimeringstiltag. Dette er et resultat af de krævende trin, der er nødvendige for en gennemtænkt datahentning fra historikken og de efterfølgende trin med forbehandling af data. Endvidere er der behov for effektive metoder, som er i stand til at håndtere data i dets naturlige struktur, svarende til den kontekst det var genereret i.

Formålet med denne afhandling er at adressere nogle af de udfordringer og vanskeligheder, som er forbundet med "genbrug" af historiske data fra en fuld-skala produktion af industrielle enzymer. Først præsenteres det afgørende og besværlige trin med at hente data fra systemerne. Forudsætningerne, som er nødvendige at forstå, diskuteres, såsom sensorernes nøjagtighed og pålidelighed, aspekter i forbindelse med den reelle målefrekvens og ikke-ækvidistante opsamlingsstrategier i dataopbevaring. Forskellige systemer kan anvendes til at ekstrahere data, og nogle kan måske introducere uønskede artefakter i de endelige analyseresultater (POSTER II). Adskillige signalbehandlingsteknikker er også kort drøftet og eksempler på deres anvendelse er præsenteret, f.eks. hvordan man kompenserer for sensorer med lav signal-støj-forhold eller håndtering af artefakter i data. En anden vigtig opgave er justering og synkronisering af procesdata. Dette er særlig vigtigt, når man ser på relationen mellem sekvenser af enhedsoperationer adskilt i tid og i endnu højere grad, når der arbejdes med (halv-)kontinuerlige processer når tidsseriedata genereres. I denne sammenhæng er potentialet af auto- og kryds-korreleret analyse og virkningen af den nødvendige de-trending af signalet udforsket i forbindelse med den kontinuerlige granulering-tørringsproces (POSTER I).

Forskningen som præsenteres i denne afhandling er primært koncentreret om ultrafiltreringstrinnet, i hvilket enzymer er oprenset og op-koncentreret. Gennemløbet af en kontinuerlig ultrafiltreringsoperation er begrænset af fænomenet ”membran-tilstopning”, hvor produktionskapaciteten - overvåget som strømningen gennem membranen eller flux - falder med tiden. Fluxen varierer betydeligt imellem hver kørsel af det samme produkt og ligeledes mellem forskellige produkter. Denne variabilitet påvirker klart produktionsplanlægningen og fører til yderligere omkostninger på grund af den hyppigere membranrengøring. Datasættet, der blev undersøgt i dette studie, var indsamlet fra målinger fra konventionelle univariate processensorer indsamlet over flere års produktion af én type af enzym mellemprodukt. Forskellige strategier for organiseringen af disse datasæt er blevet sammenlignet med varierende antal tidspunkter i datastrukturer egnet til latent variabel (LV) modellering. Det overordnede mål med dataudvindingstrinene er at komme frem til statistiske "soft models", som fanger princippet eller den latente struktur i det system der undersøges. Hvis dette fører til ny viden kan det bruges til optimering af fremtidige produktionsserier. I den primære undersøgelse blev PLS2 anvendt på data, som var reduceret til en middelværdi per produktionskørsel kombineret med nogle andre relevante variable. Det gjorde det muligt at identificere de væsentligste forskelle mellem procesvarianter af det undersøgte enzym. Strukturering af data i tre-vejs matricer er opnået ved at begrænse datasættene til deres medianlængde. Undersøgelser med LV teknikker efter strukturering af data i forhold til batche har ikke ført til nogen særlige resultater. Derfor er det blevet konkluderet, at processen kan modelleres tilstrækkelig godt, når datasættene struktureres i forhold til variablerne. De senere undersøgelser brugte denne type strukturering af data og fokuserede kun på produkter med højere koncentrationer, da disse tilfælde har det største fald i flux. Blokering i række- eller tidsdimensionen blev brugt i ARTIKEL II. Datasættet har en naturlig struktur med flere niveauer, hvor det første niveau er processens tidforløb, der er indlejret i ultrafiltreringskørslerne, som er niveau to. ”Multi-level Simultaneous Component Analysis with invariant Pattern” (MSCA-P) anvendes til at udforske dette datasæt i forbindelse med faldende flux. Vi bygger på ideen om to-niveauer og udvider modellen med et tredje niveau: »proces fremgangsmåde«. I ARTIKEL III er blokering i kolonnen eller proces markør retning blevet brugt. En multiblok PLS bryder procesvariablerne op i mindre

grupper og grupperer variablerne med samme indflydelse og karakteristika for at lette den diagnostiske procedure. Begge metoder fører til en opdeling af datastrukturerne til intuitive fortolkningsbare løsninger ved at bibeholde den naturlige struktur af de analyserede data.

Derudover er ultrafiltreringssystemet også blevet undersøgt med hensyn til produktudbytte. Potentialet af NIR-teknologi til at overvåge aktiviteten af enzymet har været genstand for en forundersøgelse præsenteret i ARTIKEL I. Den omfattede (a) vurdering af hvilken af to real-tid NIR flow celle konfigurationer som er den foretrukne til overvågning af retentat strømmen downstream til UF, og (b) om systemet kan bruges til statistisk procesovervågning og tidlig varsling / fejlfinding. Det var muligt at udvikle tilfredsstillende robuste kalibreringsmodeller til fire typer af enzymprodukter, hvor specifikke enzymaktiviteter er blevet standardiseret i en global QC parameter. Endelig fremgår det af undersøgelsen, at det mindre krævende in-line flow celle konfiguration klarede sig bedre end online-konfigurationen. Den førstnævnte virkede tilfredsstillende robust over for forskellige produkter (amylaser og proteaser) og tilhørende procesparametre såsom temperatur og proceshastighed. Denne afhandling viser, at kemometriske metoder specielt designet til to-vejs og multi-datasæt problemstillinger har stort potentiale som PAT værktøjer, da de opfylder det primære mål for PAT, nemlig at opnå en bedre procesforståelse på en hurtigere og mere intuitiv måde, især når den oprindelige datastruktur og dimension bliver bevaret.

List of publications

PAPER I:

A. Klimkiewicz, P.P. Mortensen, C.B. Zachariassen, F.W.J. van den Berg, Monitoring an enzyme purification process using on-line and in-line NIR measurements, Chemometrics and Intelligent Laboratory Systems, 132 (2014), 30–38

PAPER II:

A. Klimkiewicz, A.E. Cervera-Padrell, F.W.J. van den Berg, Multilevel Modeling for Data Mining of Bio-Industrial Processes, Chemometrics and Intelligent Laboratory Systems (2016), *Article in press*

PAPER III:

A. Klimkiewicz, A.E. Cervera-Padrell, F.W.J. van den Berg, Modeling of the Flux Decline in Continuous Ultrafiltration System with Multiblock Partial Least Squares, Industrial & Engineering Chemistry Research (2016), S*ubmitted*

POSTER I:

A. Klimkiewicz, P.P. Mortensen, C.B. Zachariassen, F.W.J. van den Berg, The value of historical data in the optimization of biomanufacturing, Scandinavian Symposium in Chemometrics (SSC13), Stockholm

POSTER II:

A. Klimkiewicz, F.W.J. van den Berg, A chemometric approach to the optimization of bio-industrial processes, Chemometrics in Analytical Chemistry (CAC-2014), Richmond, Virginia

List of abbreviations

Abbreviations

2-D	Two-dimensional
ACF, $r(k)$	Autocorrelation function
AOC	Abnormal operating conditions
BWU	Batch-wise unfolding
CCF, $r_{xy}(k)$	Cross-correlation function
CCSWA	Common component and specific weights analysis
CIP	Cleaning-in-place
CV	Cross-validation
DCS	Distributed control system
DIFF	First differencing
FE	Features extraction
iPLS	Interval PLS
LV	Latent variable
MA	Moving average
MB	Multi-block
MSCA	Multilevel simultaneous component analysis
MSCA-P	MSCA with invariant Pattern
NIR	Near infrared (spectroscopy)
NOC	Normal operating conditions
PAT	Process analytical technology
PC	Principal component
PCA	Principal component analysis
PLD	Piecewise linear de-trending
PLS	Partial least squares or projection to latent structures
PLS2	PLS with multivariate Y
QC	Quality control
r	Pearson correlation coefficient
R&D	Research and development
R^2	Squared correlation coefficient
RI	Refractive index
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross-validation
SG	Savitzky-Golay
SR	Selectivity ratio
T^2	Hotelling T^2
UF	Ultrafiltration
VCD	Volume concentration degree
VIP	Variable importance in projection
VWU	Variable-wise unfolding

Algebra notation

E	Residual matrix
F	Number of components (factors) used by the model, or number of features in FE
I, i	Number of rows or batches/runs
J	Number of variables/process tags in independent matrix **X**
K, k	Number of lags in serial correlation analysis or number of subsets in cross-validation
N, n	Number of samples, measurements or timestamps
Q	Number of variables/quality measures in dependent matrix **Y**
P	Loading matrix
q	Common loading in CCSWA
T	Score matrix
t_n	Timestamp
W	Weighing matrix or association matrix in CCSWA
X	Matrix of independent process variable parameters
x_n	Value at time n
Y	Matrix of dependent variables
λ	Weights (saliences) in CCSWA or eigenvalues in LV modeling

Table of contents

1. Introduction

Industrial enzyme production is a complex discipline where numerous critical factors have to be controlled in order to ensure a profitable outcome. The enzyme manufacturing process can conceptually be divided into three production and formulation steps (each near-autonomous 'factories'): 1) Cultivation of enzyme producing organisms, 2) Recovery of the enzymes, and 3) Formulation of enzymes into intermediate/end-products (Figure 1-1). Only a good understanding of all unit operations involved and transitions between them can ensure the final product quality and minimize unwanted product variation [1].

All the process steps in current enzyme production work well, but improvements with substantial economic and environmental impact can be achieved by making the processes work even better. The different production steps and unit operations are today run by a combination of recipe operation plus inferential control. The first part (the operating recipe) is available from engineering knowledge and equipment design (first principle or mechanistic models, e.g. mass and energy balances), while the second (inferential control) is traditional regulation based on what is easily measured [2]. At present, a considerable number of diverse measurements are thus collected throughout the different process steps, typically for dedicated univariate monitoring and closed-loop control tasks. These measurements, with a wide variety in information content and quality are stored in the historian(s), in various formats and with different sampling rates. This generates large amounts of data which are seldom used outside their direct scope (hence, real-time input to the operators and closed-loop controllers). Moreover, the combination of all the variables affecting the process plus product, and their correlations at each time interval - as well as their time correlations over the duration of the process, i.e. auto- and cross-correlations for the entire production run - are hardly ever explored [3-5]. This is a result of the demanding steps involved in thoughtful post-run data retrieval from the historian and the subsequent data pre-processing steps necessary for a more wide-scale

process data evaluation. For specifically biotech downstream processing, there is also a need for methods capable of analyzing time series data generated during continuous - or better semi-continuous - processes. This type of process shows some periodicity behavior, characterized by widely varying operation times, but is not easily classified as either batch or continuous in the classical sense. The potential of already collected full-scale production data and its related quality history is generally underemployed during the optimization efforts.

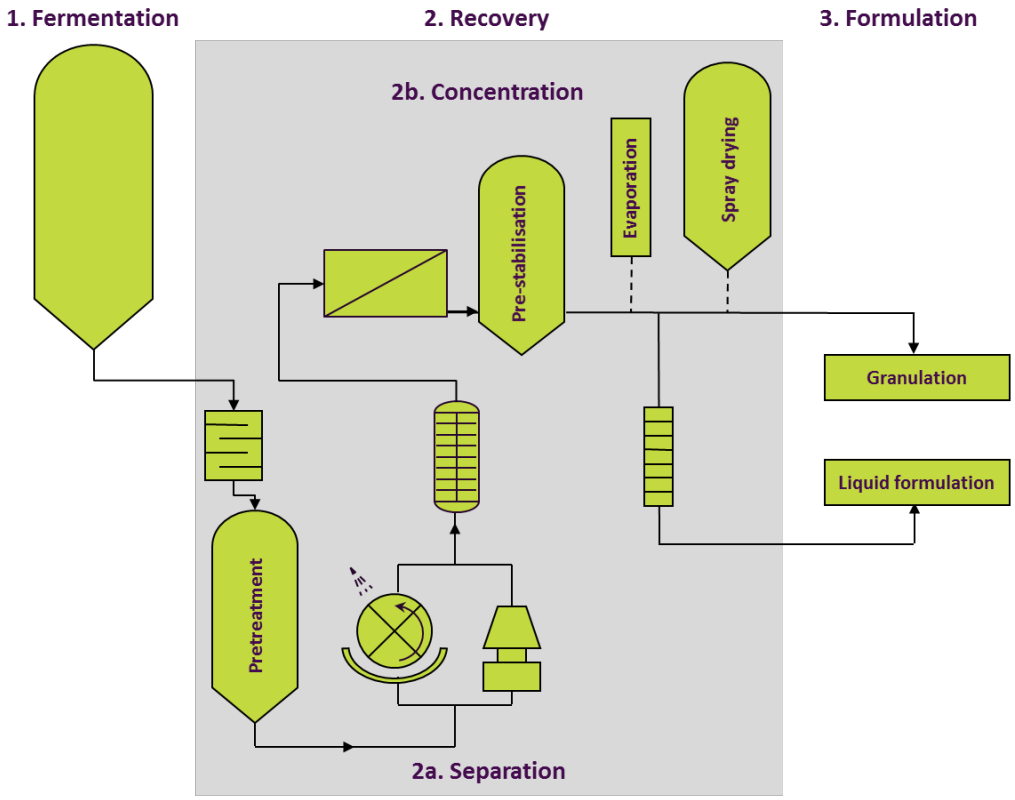


Figure 1-1 – Schematic presentation of an industrial enzyme manufacturing process.

Currently, univariate or first order principle based models are constructed and exploited to link process variables that can be measured to the desired process objectives within one unit operation or factory. Nevertheless, classical engineering strategies do not perform optimally in control and optimization of full-scale biomanufacturing steps. This can be explained by the complex multi-stage nature of the production systems which cannot be derived from mechanistic or first principle concepts [6]. The alternative, statistical model building on process data, is known as

Process Analytical Technology (PAT) [7]. An important aspect in this is Process Analytical Chemistry, the in-process (often named on-line or real-time) measurement of relevant qualitative and quantitative parameters [8]. The second facet in PAT is the multivariate statistical approach, collectively called Process Chemometrics, where functioning or behavior of a complex system under investigation is not (only) seen by mechanistic or first principle models but rather as latent or principal phenomena [9].

This Ph.D. project was driven by the expectation that combining available information stored in production historians into multisets and application of novel chemometric modeling methods will uncover extra information from historical data. It is appealing to use chemometric tools because they are capable of overcoming challenges associated with biotech applications such as multidimensionality of the dataset, a high degree of correlation between process variables, missing data and variation due to process disturbances such as noise [10].

1.1 Scientific motivations

The goal of the academic contribution is to prove that data-driven multivariate statistical methods can play an important role in process understanding, detection and diagnosis of abnormal situations and process control, structured via PAT. In particular, it is interesting to investigate the use of multiset (multiway, multiblock, multilevel) methods on full-scale production data within unit operations, on the transition between unit operations and on a plant-wide scale. Furthermore, the goal is to improve theoretical understanding from practical experience, applying existing algorithms and/or developing novel methodologies based on real-world insight and experience. An additional value of this industrial Ph.D. project is to improve education by cooperation with industry and to make research available to a wider public - already performed in the company but normally not communicated to the outside world.

1.2 Industrial motivations

This industrial Ph.D. assesses the production of bulk enzymes at Novozymes A/S by conducting multivariate statistical process analysis and optimization. Investigation of data from process sensors collected in the data historians, and interpretation of the process stages is performed in close cooperation with a multitude of Novozymes' experts. During this task, large amounts of process data were adjusted (pre-

processing, variable selection, organization in data structures) and this has led to an overall assessment of the quality of historical production data and recommendations on better practices in data compression for the future. The biotech industry can benefit from the wealth of knowledge accumulated and published over the years within the field of multivariate statistical data analysis as has been successfully applied in other industries and research areas [2,11-17]. The goal of combining multivariate statistical analysis with first principle process understanding is to turn an existing experience-based production process into a scientific-based practice. Multivariate latent variable methods are known to efficiently identify unusual operating periods and can assist in isolating the section of the plant and the group of process variables that are related to a problem [1]. They are powerful tools for focusing the attention of the operators and optimization engineers to a smaller area, allowing them to use their knowledge efficiently in diagnosing the cause of abnormal situations and thereby improve the process performance. Thus, it is anticipated that the process chemometrics/latent variable approach will aid in troubleshooting in recovery and granulation. Moreover, as an outcome of the better process understanding, it is expected that time and money spent on standard optimization work done in the company will be reduced.

1.3 Aim of the thesis

The aim of the thesis is to uncover the information and relations between unit operations hidden in already collected data by means of PAT tools. By identifying the statistical soft models from full-scale manufacturing data, it is expected to capture the principle or latent behavior of the system under investigation, which in turn can be exploited for optimization of future production runs. There is, to our knowledge, no published experience on application of multiset factor models in industrial production of enzymes. Furthermore, the aim of this project is to investigate possible techniques of data cleanup and strategies for selection of reliable sensors and to make recommendations regarding the most suitable and efficient chemometric algorithms for analysis of historical datasets from the downstream processing of enzymes.

Most of the research presented in this thesis is positioned in recovery (Figure 1-1). Specifically, the throughput of the ultrafiltration (UF) step, measured as flow through the membrane and referred to as ‘flux’, varies significantly from run to run, even for

the same product and likewise between different products. Following identification of this focus area, it is expected to draw insight from previous studies within Novozymes A/S and from the knowledge of the recovery experts to finally identify the root cause behind this variation in recovery.

1.4 Thesis outline

The introductory part of this thesis familiarizes the reader with the steps and related practical considerations involved in knowledge discovery based on historical production datasets. These include such aspects as data extraction, cleaning, appropriate dimension reducing techniques and data mining. Next, three industrial cases are discussed supported by expert knowledge and elaborated on in three scientific papers (PAPER I, II and III) and two posters (POSTER I and II). The text is organized as follows:

Chapter 2 discusses the characteristics of historical databases. Data acquisition and storage procedures are reviewed. Potential pitfalls of data compression are visualized by real examples encountered in production settings.

Chapter 3 provides an introduction to time series data and describes tools used in this thesis for analyzing such data. Methods for data synchronization and alignment are discussed. Selected pre-processing techniques are described, namely de-trending and smoothing. Lastly, the chapter treats the aspects of data validity checking and use of theoretical and process knowledge for thoughtful data preparation/selection.

Chapter 4 starts with an explanation on how most multivariate chemometric methods are applied today in the form of traditional two-way data analysis. Next, differences in the modeling of a continuous vs. batch processes are discussed. This is accompanied by the description of data arrangements and scaling tactics specific to the above processes. Finally, this chapter ends with the description of methods tailored for particular types of processes.

Chapter 5 contains the introduction to three industrial cases which were approached with the previously described ideas. The chapter contains general information about the sequences and unit operations involved in downstream processing of industrial enzymes. The ultrafiltration system has been given the most attention as this processing step has been investigated both in terms of product yield (enzyme

activity, PAPER I) and in terms of capacity optimization (PAPER II and III, POSTER II). The second part provides introduction to the optimization study anchored in the granulation factory (POSTER I).

Chapter 6 gives concluding remarks and formulates future perspectives on the matters investigated in this dissertation work.

2. The nature of process data

Data management for food-, chemical-, pharmaceutical- and bio-processing is very complex as signals are simultaneously generated by multiple devices, and information needs to be available promptly in a common format. Databases which log and store the time-based electronic records from production facilities are called operational historians (as supplied by e.g. OSIsoft's PI, Honeywell's PHD, GE's Proficy). Process data can automatically be collected from many different sources (control systems, process outputs, physical parameters, manual entries, calculations, laboratory analysis and/or custom software). Users can later access this information using a common set of tools (e.g. MS Excel, a web browser, display interfaces in the operator rooms). The collected signals vary greatly in format and sampling frequency which can range from fraction of a second to days, and this is a first hurdle to take in data reuse. The exploitation of historical databases is an important first step towards process understanding and improvement [1]. The tendency is to store large amounts of data but routinely use only a few pieces of selected information. This is a result of the demanding steps required in thoughtful data retrieval from the historian and the subsequent data pre-processing steps. Consequently, most data is seldom utilized outside their direct scope and a powerful option to describe the 'overall process signature' is thus currently under-employed in biomanufacturing. The aim of uncovering knowledge hidden within historical production data can be accomplished only if two preconditions are met. First, one must be certain that data is reliable and relevant. Second, datasets consisting of different types of process runs need to be assembled.

2.1 Process signals

Bioprocess plants keep an abundance of electronic records collected during different manufacturing steps all the way from the seed culture, through the full-scale bioreactors to downstream processing and formulation steps. Data can be related to material input (quantity, suppliers, results of quality analysis conducted in the laboratories, etc.), process outputs (fermentation titer, cell density, product

concentration, etc.), control actions (acid dosing, flow rate, etc.) as well as physical parameters (conductivity, temperature, etc.).

In general process instrumentation can be split into two broad categories: sensors and analyzers [18]. Sensors are compact, self-contained devices with most of the supporting utilities on board (e.g. turbidity, conductivity and pH-sensors). Analyzers are bulky instruments demanding various external utilities (power, air conditioning, etc.), and routing of cabling or fiber optics to the flow cell or probe (for instance spectroscopic or particle size distribution systems). A further distinction is that process data can be univariate, such as a pH, temperature or pressure readings, or multivariate as provided by most spectroscopic tools. The frequency of the acquisition of data generated during the manufacturing process can be categorized into discrete, intermittent or continuous [10]. In the production environment, each item being logged is called a 'process tag'. Most of the process data is acquired continuously on-line e.g. for instantaneous monitoring in the control room or local closed-loop control strategies. However, not all data acquired at designated intervals is stored. Only the intermittent data which passed the compression testing is archived in operational historians. Among the process parameters one can find also binary data, for instance for valves opening and closing which can only take 'ON/OFF' values, or descriptive tags – which inform on operational phase of the process or store information on alarms. Some parameters are measured at-line next to the production line or in dedicated laboratories located near production. A result of these discrete analyzes usually ends up in paper format on the production control cards (and ideally in digital format after some time). Finally, many key parameters related to raw material, final or intermediate product quality (such as viable cell count, activity, concentration of impurities, etc.) require time-consuming off-line analyzes in the central lab off-site or outsourcing the service to a contract lab. This information, usually required for the final product release or linked to the production trials, is stored in Laboratory Information Management System (LIMS) and Enterprise Resource Planning (ERP) software such as SAP. Additionally, technical information, for instance, regarding dates of replacement of particular part of the equipment, or other extra information related to production trials may be stored in files on the shared drives or in the form of paper documents in the archives. Consequently, data is stored in so many different formats and places within a plant that it is not a trivial task to combine it all, and in particular not in a timely manner

which could offer a full picture of the interrelations and importance in process control and product quality [18].

This Ph.D. project predominantly uses data digitally stored in the process historian. Therefore a detailed explanation on related data archiving and retrieval is provided in the following section.

2.2 Data acquisition, compression and reconstruction

It is relevant to comprehend that, in order to save storage space in the historian, not all measured data points are stored. The compression is also driven by other motives such as cutting down the network traffic and improving performance when retrieving the data [19]. What is more, providing that the tolerance settings match the precision of an instrument, noisy, unreliable information should be automatically discarded before archiving. Eventually, it is expected, that the trends reconstructed from archived data preserve the fidelity of the raw data signal. This is however not always the case, and e.g. MacGregor and Kourti [20] warned already in 1995 that the univariate compression tactics can destroy the multivariate features in the process data.

On-line data acquired usually undergoes filtering/compression at one or several of the following stages:

Instrument → Distributed Control System (DCS) → Interface Node → Historian

There are no general guidelines regarding at which of the above steps compression should be implemented. Even within one business it can vary from one site to another and usually depends from the programmer administrating a particular system. As a starting point, a quick overview of the true logging frequency for a tag can be obtained by downloading all archived data for a certain period and by referring the data count to the requested time range. As a consequence of individual compression settings of different tags, archived data is heterogeneous with respect to time scales. For instance, Table 2-1 offers an overview of the compression information related to one of the ultrafiltration units in the recovery factory at Novozymes A/S in Kalundborg. Here, conversion of all process tags (except 'Operation') takes place in the DCS system, but not in the historian.

Table 2-1 - Overview of compression parameters of selected process tags collecting information during ultrafiltration.

Selected Tags			Compression summary	Historian compression	DCS compression parameters		
Tag mask (UF-X_*)	Unit	Description	Logging frequency per max log. interval	Compressing on historian server	Dead band (eng. unit)	Pooling interval (-)	Max logging interval (-)
FFT01	$m^3 \cdot h^{-1}$	permeate flow	147.9	OFF	0.06	0.001	0.2
FFT02	$m^3 \cdot h^{-1}$	dilution water flow	6.5	OFF	0.03	0.003	0.2
FFV01	%	feed reg. valve	173.5	OFF	1	0.001	0.2
FJT01	%	power of circ. pump	5.6	OFF	1.3	0.001	0.2
FNT01	R	dry matter (1)	3.6	OFF	0.2	0.001	0.2
FTT01	°C	temp. recirculation loop (1)	3.2	OFF	1	0.001	0.2
FTT02	°C	temp. recirculation loop (2)	2.9	OFF	1	0.001	0.2
YCT01	$mS \cdot cm^{-1}$	retentate conductivity	106.3	OFF	0.1	0.001	0.2
YFT03	$m^3 \cdot h^{-1}$	UF retentate flow	115.7	OFF	0.01	0.002	0.2
YFV03	%	reg. valve UF retentate	59.2	OFF	1	0.003	0.2
YNT01	R	dry matter (2)	1.3	OFF	1.6	0.003	0.2
YTT01	°C	temp. feed tank	1.3	OFF	1	0.003	0.2
YTT02	°C	temp. retentate	68.3	OFF	1	0.001	0.2
ZAT01	-	pH after permeate tank	2	OFF	0.14	0.003	0.2
Operation		operation mode	6	ON			

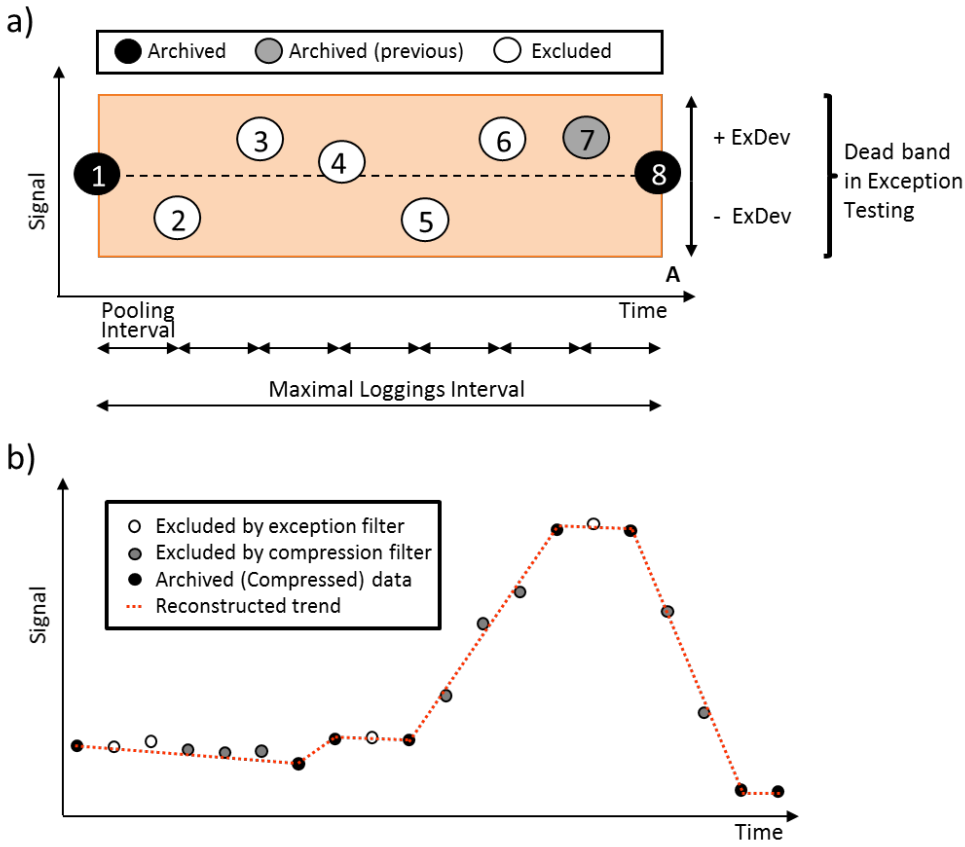


Figure 2-1 – Effects of data compression; (a) the dead band in exception testing; (b) effect of popular filtering techniques; pictures loosely based on [19].

The key term when it comes to any compression type is the 'dead band' (see Figure 2-1a). It is a certain tolerance given to each variable, usually based on the range of this variable throughout the process and/or the instrument precision. Explicitly, it defines how much a new value has to differ from the previously logged (old) value before it is considered significantly different. Generally, the dead bands should be slightly narrower than the instrument precision. If a new recording is outside the dead band, it merits storing the new value (together with its timestamp). The associated parameters which also need to be set are the 'Pooling Interval' and the 'Maximal Loggings Interval'. The former is the minimum time distance between the evaluated data points. The latter is the maximum time span between logged values.

There are two popular types of tests used as filters:

1) 'Exception' reporting (Figure 2-1a) takes place on the interface node before the value is sent to the historian server. The newly pooled value passes the exception test when the difference between this value and the last archived value is greater than 'ExDev' (falls outside the dead band). That value and the previous value are reported to represent adequately the actual behavior of that process parameter (or more precisely: the process trend). Alternatively, the value is logged if the difference between the times of the new value and the last archived value is greater than 'Maximal Loggings Interval' (which is the case for the situation presented in Figure 2-1a). The aim of exception reporting is to reduce the communication burden between the server and the interface node by filtering out noise [21].

2) 'Compression' testing takes place on the historian server subsystem before data is sent to the archive. The idea here is not to store an event that can essentially be recreated by interpolating from neighboring events. This test uses a so-called 'swinging door' algorithm which allows for a dead band to have a slope [19]. The maximum and minimum slope is estimated based on: the most recent archived value, the current value that passed exception testing and the compression deviation settings. The reference slope is calculated based on the incoming value (next value that passed exception) and compared to the maximum and minimum slopes. The rule applies here that the maximal and minimal slopes can only get narrower when subsequent values are evaluated vs. the last archived value. If the reference slope falls outside the calculated limits, the current value becomes the next archived value. In the end, between any two archived values one can draw a parallelogram (a dead band defined by compression deviation settings) that holds all events that happen between these two archived values.

An effect of both filters is shown in Figure 2-1b. In relation to these procedures, e.g. Kourti [11] confronts how, especially if setting the tolerance limits too wide, it can lead to the introduction of spurious correlation between variables.

In Table 2-1, only the tag 'Operation', which stores information on the current operation mode of the unit, uses the compression option on the historian server. The dead band parameter 'Compression Deviation' (not shown) is in this case set to 0. This means that the successive identical values (or values aligning perfectly along a

sloping line) are not archived. For example, if the last logged entry is 'Startup' and it is followed by another 'Startup' entry, the second entry will not be added. If the operational sequence changes e.g. to 'Filtration', it will be recorded; otherwise entries are logged according to the provided 'Maximum Loggings Interval' setting.

Various strategies can be used to reconstruct the trends from preserved data out of which the following three are the most commonly selected:

1) 'Compressed Data' returns all of the values that were logged during the specified time range. This function returns the archived values which are stored in the historian and which passed the exception and compression testing.

2) 'Sampled data' reconstruction is the most convenient to use. The user defines the time span and the sampling (or more accurate signal reconstruction) frequency, and the function retrieves values evenly spaced in this time span. These values are interpolated from the values stored in the archives. The user may miss maxima and/or minima in overall trends.

3) 'Archived value' in retrieving mode set to 'previous' returns the last logged value instead of interpolation. It requires that the user specifies a timestamp array/vector, contrary to the two previously described functions which require just the start- and end-point of the investigated time range.

Effects of the above-described functions are depicted in Figure 2-2 for two process tags which record the same physical property but differ in tolerance settings where signal reconstruction functions are compared for two UF runs.

The compression parameters of tag 'YNT01' are set erroneously. The dead band is set to 1.6 R while the dry matter changes equivalent to 0.2 R should have been captured as a minimum. With the current settings, the corresponding data are logged primarily according to maximal loggings limit. The tag 'FNT01' has its dead band set to 0.2 R which fits the purpose better. Still, the resulting logging frequency is relatively low in comparison to other process parameters. In the case of run A, the important fluctuations in the dry matter happening between times 0.4 - 0.8 captured by the first sensor (Figure 2-2a) are missed in case of the second one (Figure 2-2b). Interpolation with the 'Sampled' function works well for visualization of the process trends, but only if the tolerance limits are set appropriately and when the sampling

frequency is sufficiently high. In the case of the tag 'FNT01' the 'Archived' retrieval function provides almost equally good or sometimes even better representation of the true trends of this variable. This is because the ultrafiltration operation is controlled based on the set-point in concentration ratio and sudden jolts and step changes may occur. For instance, in the case of run B and tag 'YNT01', it can be seen that it is no longer possible to retrieve true trends which involved sudden jolts (Figure 2-2c vs. d). Instead, broad peaks or artificial slopes are introduced during reconstruction.

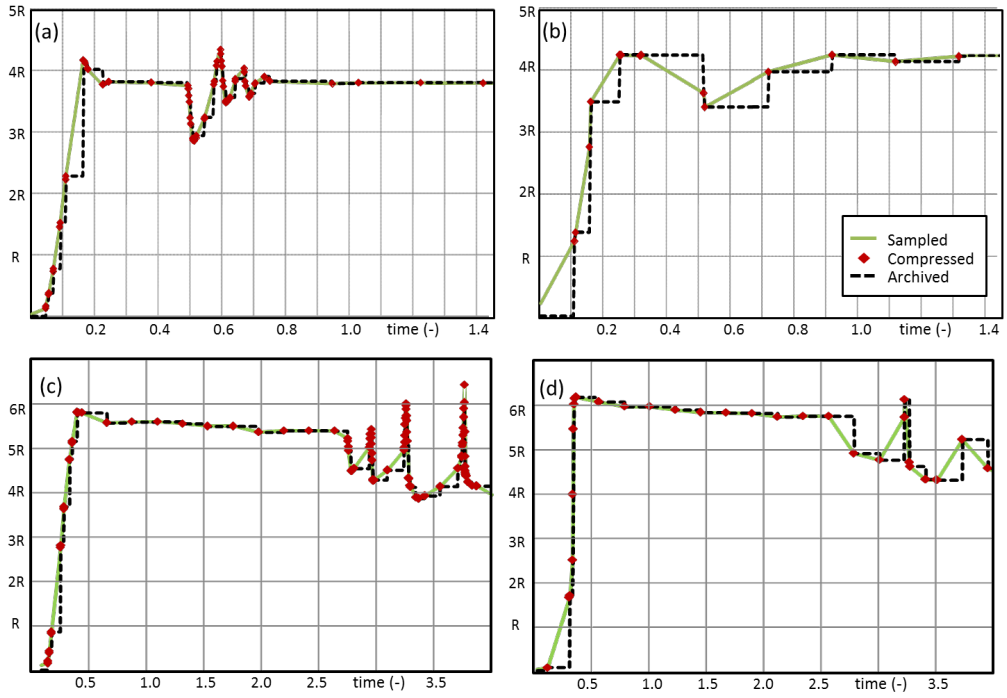


Figure 2-2 - On-line dry matter measured downstream to the UF – comparison between the results from three different data reconstructing methods; (a, c) tag: 'FNT01' (located on the last UF recirculation loop); (b, d) tag: 'YNT01' (located in the retentate stream); (a, b) run A; (c, d) run B; compare with compression parameters summarized in Table 2-1.

Consequently, it is very important for the practitioner to examine the true trajectories of the variables and decide on the appropriate compression settings and reconstruction function with a suitable data storage frequency that best preserves the process behavior [13]. It is a known issue that the multivariate nature of the data is not preserved during the conventional, univariate data compression and processing

[11,13,22]. It has been suggested that most of these problems could be solved by decreasing the tolerance limits [11,13], and at a minimum a periodic evaluation of data storage strategies is advisable. Alternatively, MacGregor and Kourti [20] proposed to store the scores of the first predefined latent variables and the loading matrix. From this, the original variables can always be reconstructed as long as no special events, which are not predicted by the model, occur.

To summarize, first of all it is important to set filtering parameters so that adequate information is captured to match the purpose of the future use of the sensor data. Making the right choice is handicapped by two aspects: ownership of the signal and incentives for optimization. As stated, most signals are generated for local, closed-loop control or alarm detection. The primary objectives of the process tags might thus be very different of that in long-term, history based process investigations. A related difficulty is that by convention the historian is meant for archiving, where simple objectives are to preserve the main operations while keeping network traffic low. This is not necessarily compatible with special occurrences such as the spikes observed in Figure 2-2c. This leads to a classical chicken-and-egg situation where extraordinary incidences that might require a special action are not preserved by the data historian, and only after they have been identified they are stored. The second step, retrieval mode, needs to be chosen wisely to reconstruct the true, relevant process trajectory. If these two aspects are not defined correctly, one can miss important process fluctuations as was shown in the dry matter example. The appropriate settings should be agreed among process engineers and operators, consulted with statisticians and communicated to the programming personnel. Unfortunately it is not common practice yet to pay attention to this before a specific challenge has arisen. The reader interested in knowing more about the aspects of data archiving and extraction is recommended to consult the existing literature, reports and webinars [11,13,19,21].

2.3 Assembling the data

Prior to any big data mining challenge it is essential to organize thoughtfully the information into data structure that assemble all relevant processing and quality parameters for each lot. In each of the key factories involved in enzyme production (Figure 1-1) a specific volume of the product is assigned a batch number. For instance, in production of the enzyme granulate, one needs to align the

corresponding recovery and fermentation batches. Moreover, one recovery batch can consist of two or more fermentation batches and vice versa (a 'split batch'). Then again, the granulation factory operates in 'campaigns' which can consume several recovery batches and produces several granulation batches. Even though in modern production good traceability is always in place, a collection of the data representation for more than one final batch at the time is not a trivial task as there are no automated ways of doing this. In practice, alignment of the information from downstream processing is even more complex when processes run semi-continuously. These types of processes show some cyclical behavior, characterized by distinctly varying operation times. As a consequence, unit operations are shifted with respect to one another and vary considerably in duration. Therefore, the time series corresponding to different unit operations not only need to be matched but, preferably, also aligned. Other complications in data assembling related to a semi-continuous production can arise from following situations:

1) One operation can run on different equipment from time to time, for instance normally assigned to another production line.
2) Some processing stages use several units in parallel but in different combinations, for example, when two units perform the operation the third one is off or being washed or loaded; afterwards, all three are in use again for some time until unit one is shut down.
3) By-passing - when one concentration step is omitted and another stage takes over.
4) Unplanned stops, for instance, due to membrane fouling; now all upstream operations need to wait until this unit is cleaned and ready.
5) Production trials involving new or modified process stages.

The above situations are primarily dictated by the optimal use of the production capacity and obviously make the thoughtful data extraction very demanding.

During the investigation of historical data, it may already be known that some stages of the continuous recovery process are sufficiently well described by one single value, for instance, the mean dose of flocculation chemical in pretreatment. This can be done if adequate process experience exists. However, it might happen that step changes occur in the usually stable parameters due to production trials. Additionally, one might consider using information related to the equipment wear, team of

operators being on a shift, or information gathered in manufacturing recipe revisions which is not covered by other process or quality parameters. Naturally, analyzing at the same time process phases described by static parameters and dynamic phases better described by time series needs special attention. Aspects related to this are further discussed in Paragraphs 3.5 ('Data reduction') and 4.6 ('Modeling batch vs. continuous processes').

Another important subject is how to define a batch or other section of the production corresponding to one entity during our data mining. This can differ between examined unit operations, factories or parts of the process and it should be primarily determined by the aim of the analysis. E.g. the definition of a batch in production systems is not always in line with the purpose of our studies. In production systems, 'batch' is an administration number linked to a 'process order'/'lot of material'. Hence, we often talk about 'batch' even in continuous manufacturing.

The task of selecting the correct segment of the process is shown based on the UF capacity study (POSTER II). In the historian, data assigned to one administrative batch number 'B-010' can be extracted using a previously determined start and end time of this particular batch at this exact unit (here called 'UF-X'). All data obtained in this procedure is represented by a black line in Figure 2-3.

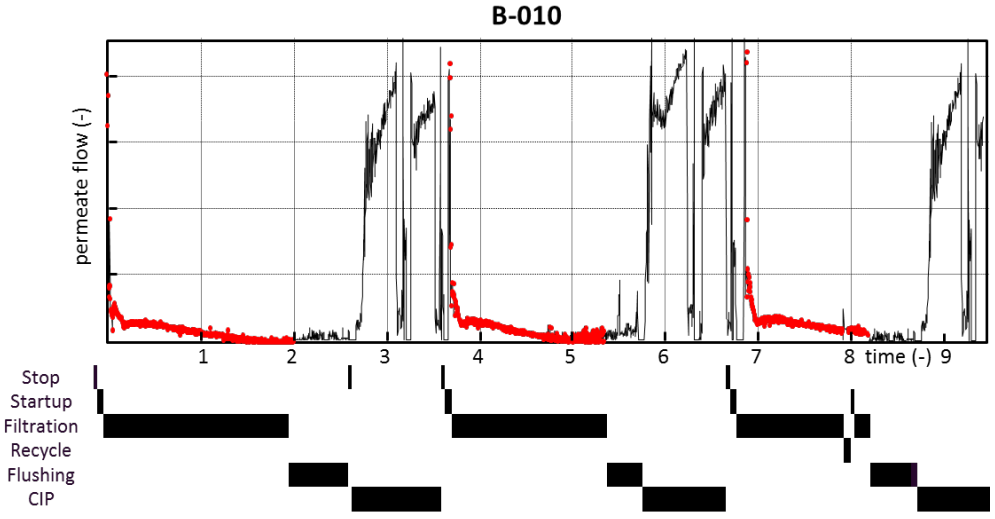


Figure 2-3 - Records logged under the tag 'UF-X_FTX' - 'permeate flow' - for all operational sequences of batch number 'B-010' processed on the unit UF-X; data assigned to 'Filtration' is marked in red; time and permeate flow is expressed in arbitrary units.

The first important step is to dissect only data relevant to the problem under study. The following operation sequences are available on 'UF-X', which are called by 'UF-X_Operation' tag:

1) Stop 2) Startup 3) Filtration 4) Recycle 5) Flushing 6) Cleaning-in-Place (CIP)

The purpose of the study used as an example was to investigate the membrane fouling problem occurring during a steady-state phase of the UF process. Therefore, only the data assigned to '3) Filtration' are of interest. It is possible to specify a filter in the function used for data extraction so that only data labeled '3) Filtration' is delivered (marked by red in Figure 2-3). This automated approach has two obvious drawbacks. First of all, one administrative batch can include more than one entity corresponding to the following theoretical definition:

'Continuous-mode ultrafiltration process = type of operation where the feed is continuously supplied to the membrane plant' [23]

In the case of batch 'B-010' shown in Figure 2-3, it was necessary to stop the UF (discontinue the feed) and start a CIP sequence, twice. It means that there were three entities complying with the definition presented above, not one. CIP is initiated either by an unacceptably low flux (one parameter in a more complex economic optimization) or because the order has been finished (the true end of a batch). The task of the UF capacity project was to evaluate a flux decline which is a result of membrane fouling. As cleaning recovers membrane capacity, entities in the above analysis should correspond to CIP-separated filtration sequences; let us call them 'runs'. Subsequently, there is a need for an additional data cleaning step. If there is an extra CIP under the same administrative batch number, the data that follows is considered a new run and named as the original batch number with an index (subscript 1, 2, 3, … etc. depending on the number of unplanned CIP's). If a discontinuity in '3) Filtration' is short (called '4) Recycle') the corresponding data is excluded but the data after the break is assigned to the same run.

The second difficulty which also needs to be addressed in the above UF example is that data assigned to '3) Filtration' is not explicitly a steady-state process. Even though there is a distinct operation sequence called '2) Startup' set in the DCS, it covers only the first minutes when the unit is pressurized following the CIP phase. It is hard to describe the startup phase in terms that cover all cases. In the UF capacity

study, it was assumed that the true startup finishes when the dry matter set-point is reached in the retentate stream which, from this point in time is redirected further downstream. This regular pattern of an initial increase in dry matter at the startup could precisely be seen in the examples presented in Figure 2-2. At the same time, a fast flux/permeate flow decline can be observed. The final data obtained applying the above assumptions, which corresponds to the example shown in Figure 2-3, are represented in red in Figure 2-4.

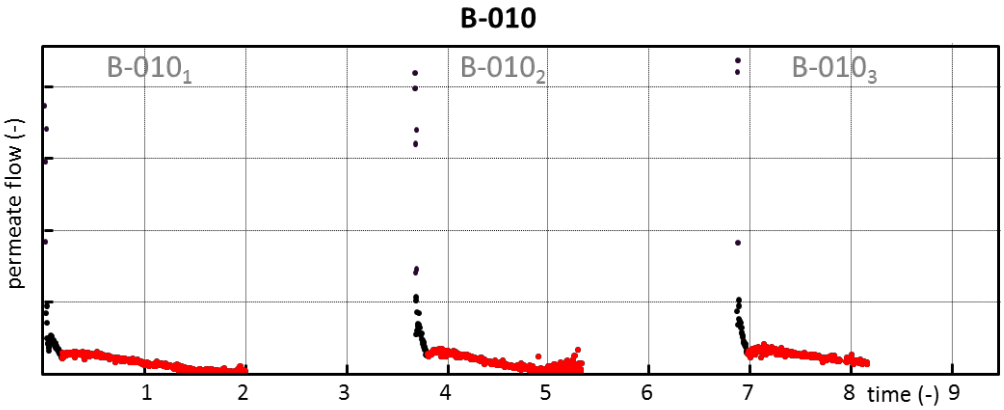


Figure 2-4- Final steady-state filtration data (marked in red) corresponding to Figure 2-3.

To summarize, the administrative batch number from the production systems is frequently not sufficiently unique for the problem under investigation. This, unfortunately, leads to a lot of manual work during data assembling, thus hampering the use of historical data evaluation. A second example on the selection of information related to a granulation case study is presented in POSTER I. Here one entity corresponds to one production 'campaign' of the specific enzyme granulate. Downtime data irrelevant for the process dynamics was excluded based on the operational tag 'In-flow Dosing Weight'. Artifacts introduced by flow-stoppages are present up till 30 min post-downtime, and this adjacent data is also taken out of the analysis.

3. Data pretreatment

3.1 Working with time series

A time series in the context of data analysis is the realization of a stochastic process in the time domain. It is a sequence of time and value observation pairs (t_n, x_n) with strictly increasing time. Production data stored in process historian (usually) form an unevenly spaced time series due to the measurements actually being performed in a random fashion or because of to the data storage strategy discussed in Chapter 2. Most of the core theory for time series analysis had been developed for equally spaced or equidistant data, due to limitations in the computing resources at the time of development, although methods of analysis for non-equidistantly spaced data are also available [24,25]. Still, the most popular approach is to transform unevenly spaced data into equally spaced data by some form of interpolation. This is also the approach followed in this thesis work since sampling intervals of the archived production data are not uniform and vary depending on the compression settings within one process tag as well as between different tags. During the data acquisition step/historian query, the sampling interval is made equal using first-order linear interpolation to a constant value for all process parameters.

Time series analysis comprises statistical methods for analyzing and modeling of an ordered sequence of observations. There are many reasons to study time series, for instance: 1) characterization of the signal or time series; 2) identification of the phenomena governing the series and modeling of the system (the 'system dynamics'); 3) prediction of the future values 4) optimal control of a system; 5) intervention analysis [26,27]. It is extensively used in any domain which involves temporal measurements such as econometrics, engineering sciences, biological studies, astronomy, weather forecasting and many more. During this thesis work, tools from time series analysis have been applied to get better insight into system dynamics and move towards variance reduction. In control, the aim is to diminish the process fluctuations and manufacture a product within certain specifications.

Process fluctuations are an outcome of the dynamics of a system. They can be modeled either 1) directly, using an impulse or a step response or 2) indirectly, by analyzing the characteristics of the observed fluctuations [28]. The first approach is less favored as it requires that the process is artificially perturbed. The second one is preferred as it can also be used when the system is under regular control. For the second approach one must however always keep in mind that the observed dynamics is the system plus control (or response) dynamiscs unless special investigations are conducted e.g. by superimposing a psuedo random binary sequence on top of the actaual feedback signal. In the remaining part of this chapter the focus is on the dynamics of the process signals in the time domain, although, it is also possible to analyze fluctuations in the frequency domain [26,29,30].

The Pearson correlation coefficient (r) is a product-moment correlation coefficient, which measures the strength of the linear relationship between a pair of variables [31]. It is obtained when the covariance of two variables is divided by their standard deviations; hence, it is independent of the measurement units. If the correlated data sets are not analyzed at one point in time but sequentially through time, then the correlation between them is called serial correlation [27,32] or lagged correlation [29]. The purpose of serial correlation analysis is to compare signals and to calculate their relationship with regard to a change in time or distance. A serial correlation where the second set is a repeat of the first is called an autocorrelation function (ACF). If the second set is another variable, then the relation is referred to as a cross-correlation function (CCF). Serial correlation analysis is a suitable device to define how well the process fluctuations are predictable. ACF and CCF are also the key tools for identification of the right complexity or order of time series model. Serial correlation functions assume that signals are stationary. This means that the mean and the variance of the process are constant over the sampled time period and that correlation between successive observations depends only on the time lag (k). Obtaining a stationary signal usually involves 1) cleanup, where data irrelevant to the process dynamics is excluded (e.g. measurement spikes); 2) de-trending, where the large-scale or 'slow' variation dictated by the set-point value changes - such as production speed - is removed. The first part, the data cleanup, has been described in the Chapter 2. It involves identification of the steady-state production regions in the data, removal of downtime due to stoppages, exclusion of system startups which are governed by distinctly different dynamics, or other artifacts in the data. Prior to de-

trending, signals can be also subjected to various types of filters to reduce the contribution of high-frequency signal components (instrument noise, local turbulent flow behavior, etc.). De-trending, on the other hand, is a form of high-pass filtering, which often forms the most critical step in serial correlation analysis.

3.2 Serial correlation functions

Not all signals are fit for the lagged correlation analysis. If the signal shows no variation, either due to strict control or a limitation in the recording system, its autocorrelation over time is virtually equal to one. This kind of signals can be quickly identified either visually or by checking the frequency of its recording in the process historian. It is always a good idea to consult the actual logging frequency of the signal, at the sensor side, as there is no point to investigate the lagged correlation on a minute scale if the reading is logged just once per hour. It is also important to consider the signal validity and the artifacts which may arise as a consequence of the function selected in the historian query.

3.2.1 Autocorrelation function

Autocorrelation stands for the correlation of a time series with its own values in the past and future. ACF for time difference (lag, k) is given by:

$$r(k) = \frac{1}{N-k}\sum_{t=1}^{N-k}\frac{(x(t)-\bar{x})(x(t+k)-\bar{x})}{\sqrt{s_x^2}} \qquad (3\text{-}1)$$

N: total number of observations

$x(t)$: value at time t

$\bar{x}$: mean of the $N-k$ observations

s_x^2: variance of the $N-k$ observations

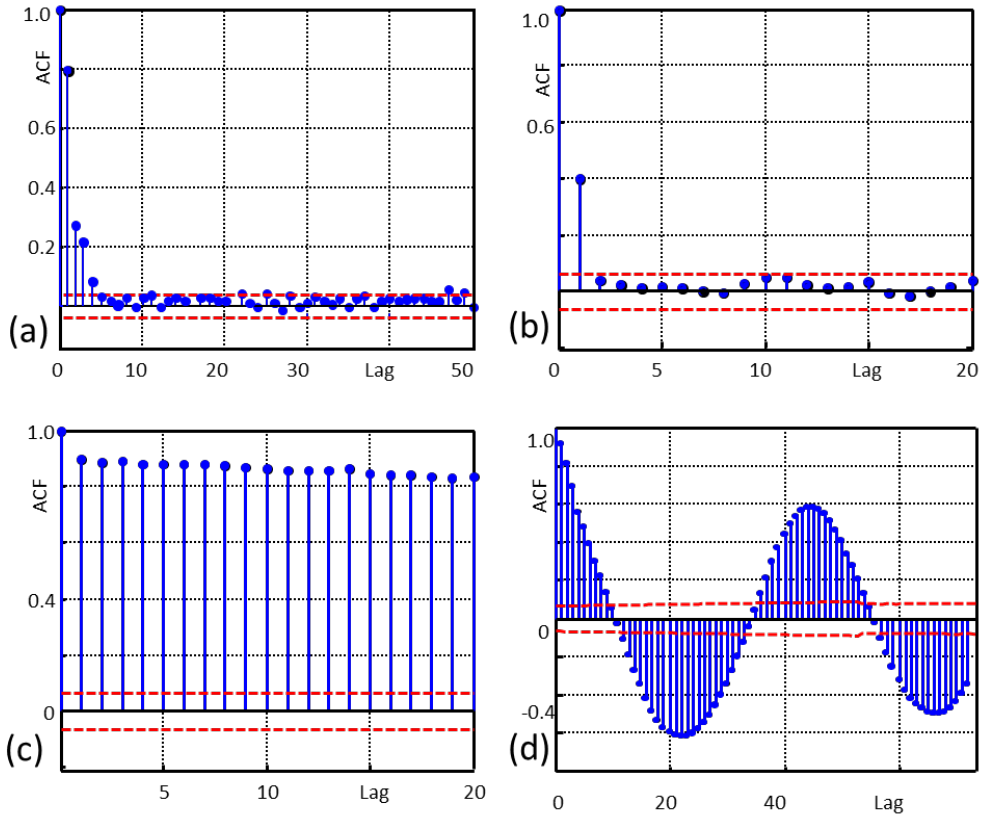


Figure 3-1 – Examples of autocorrelograms obtained for (a) some stationary signal where autocorrelation between values distanced by five lags and more is insignificant (b) a moving-average (MA(1)) process; (c) a non-stationary signal; (d) a time series with a seasonal component; red dashed lines indicate the approximate 99% confidence bands which are equal to ±3 times square root of the approximate variance of autocorrelation equal to 1/N.

For successive lags $k = 1, 2, \ldots, K$ this give $r(1)$, $r(2)$, …, $r(K)$. These values constitute the autocorrelation function. A rule of thumb in calculation of ACF (and CCF as discussed later) is that the maximal lag for which the coefficient is calculated should not exceed one-fourth of the length of the examined time series. For $k = 0$, $r(k)$ is equal to 1. The function takes values from -1, to +1 and is symmetrical toward lag zero. A plot of the values of the autocorrelation coefficients arranged as a function of lags is called an autocorrelogram (Figure 3-1). Under certain assumptions, the statistical significance of a correlation coefficient depends just on the sample size or signal length, that is on a number of independent observations [29].

In general, the ACF is expected to decay quickly to zero as is the case for the process shown in Figure 3-1a and b. Cyclic patterns in the time series lead to periodic autocorrelograms as depicted in Figure 3-1d. Since noise is uncorrelated for the large k values, periodicities are easier to detect from autocorrelograms than from the original process data [28]. Other trends in the ACF indicate that the data might first need to be submitted to filtering to correct for e.g. an unstable variance or trends (Figure 3-1a). By applying those filters and subsequent decomposition of the data, the main events governing the process are identified. Positive autocorrelation indicates a tendency of a system to persist in the same state through time [29] while negative correlations may indicate feedback in the system.

Partial ACF is another useful tool which offers more information on the correlation structure in the data. The partial ACF is the autocorrelation at lag k after the autocorrelation is first removed by an autoregressive, AR($k-1$) model [29]. Hence, it measures the correlation between signals that are shifted k lags without the effect of the intermediate values.

3.2.2 Cross-correlation function

If two correlated time series, x and y, are shifted so that they are offset in time as demonstrated in Figure 3-2a, simple correlation analysis may be misleading and strategies using CCF are more appropriate [33]. The cross-correlation function of two time series is a product-moment correlation as a function of time offset, k, between the time series, given by:

$$r_{xy}(k) = \frac{1}{N-k}\sum_{t=1}^{N-k}\frac{(x(t)-\bar{x})(y(t+k)-\bar{y})}{\sqrt{s_x^2 s_y^2}}\ , \text{ for } k = 0, 1, 2, \dots, (N-1) \qquad (3\text{-}2a)$$

$$r_{xy}(k) = \frac{1}{N+k}\sum_{t=1-k}^{N}\frac{(x(t)-\bar{x})(y(t+k)-\bar{y})}{\sqrt{s_x^2 s_y^2}}, \text{ for } k = -1, -2, \dots, -(N-1) \qquad (3\text{-}2b)$$

In contrast to the ACF, the CCF is asymmetrical which brings the need for two parts of the Equation (3-2). The CCF can be described in terms of 'lead' and 'lag' relationships [29]. The first part of the Equation (3-2a) applies to y shifted forward relative to x. With this direction of shift, x is said to lead y. This is the same as saying that y lags x. The second part of the Equation (3-2b) describes the reverse situation. The maximum cross-correlation should be achieved for properly shifted signals.

Figure 3-2 presents an example where the maximum cross-correlation is achieved at lag 5 meaning that x leads y by 5 lags.

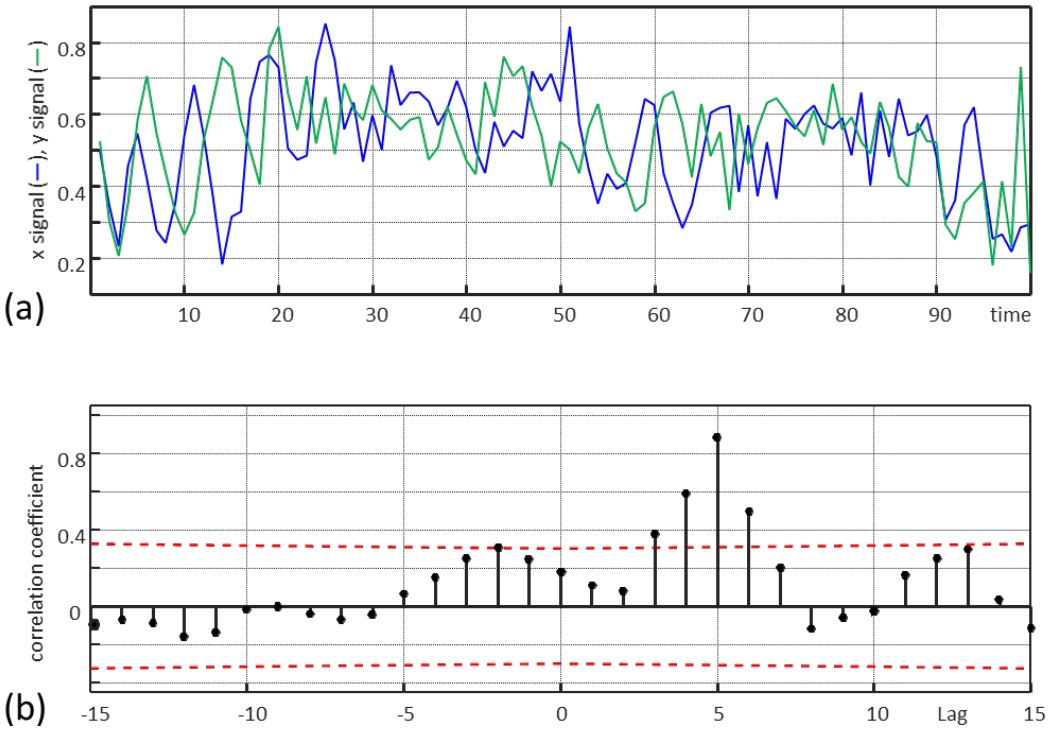


Figure 3-2 – Example of two signals which have a delayed relationship in time (a); cross-correlation diagram expressing the lagged relationship between x and y with 99% confidence bands (red dashed lines).

3.3 Signal filtering

3.3.1 Smoothing

The main aim of this pre-processing step is to eliminate noise that may be present in process measurements due to instrument limitations and sampling artifacts. In other words, smoothing is helpful with wild patterns in the data [32]. Several different filters working in the time domain have been checked throughout this Ph.D. project. It has to be remembered that filtering is by definition destructive. If used thoughtlessly it can 'iron out' valuable information or introduce artifacts in the signals. Therefore, the effect of each filter has always been visually inspected, and some examples will be discussed in this section. First off, not all signals need to be subjected to filtering. For instance, well controlled set-point-like values and slow-

moving signals might be ready for analysis in their original form. Besides, if the compression settings are set thoughtfully then only the relevant data is returned during data acquisition. In practice, when the signal is logged with a frequency lower than ten times per hour there is little chance for improvement through filtering. On the other hand, if the process tag is logged several times a minute or if the measurement system is primitive then the signal can benefit from a low pass filtering as was e.g. applied in POSTER I. The noisy data in this granulation process study has been first subjected to a 'spike filtering' which removed the most apparent outlying points from the data based on three input parameters: the time vector, a symmetric window width and a multiplier for the standard deviation. The window moves sequentially over the signal and compares the value at its center with the average value over the entire window. If the value at the center exceeds the standard deviation over the window times the multiplier then the center point is replaced by the average value. The filter takes into consideration the timestamp vector to evaluate if the compared data points originate from neighboring time points so as not to compare data which was originally separated by production break (e.g. by excluded as downtime period). The same precaution has been in place in other filters applied to process data. The second filtering step in the granulation study was the 'box-car smoothing'. This type of pre-processing produces a time series in which the importance of the spectral components at high frequencies is diminished [29]. The simplest form of smoothing consists of using a moving average (MA) where a window of an odd length is defined. The central point a window is replaced by the average value over that window [34]. When used correctly the filter can e.g. compensate for sensors with low signal to noise ratio. The smoothing or convolution function employed in MA is a simple block function. It is also possible to draw a more complex convolution functions which offer a similar or better signal-to-noise ratio with less deformation of the basic deterministic signal. The most popular option applied for this type of smoothing is the Savitzky-Golay (SG) algorithm [35]. SG fits reduced-order polynomials to all points within a moving window for estimation of the value at the center point. One of the advantages of using polynomial filters is that smoothing and derivation can be done simultaneously. SG has been used for pre-processing in PAPER I to remove physical phenomena in spectra and, as a consequence, improve the model performance. Alternative options to SG are smoothing splines. They are piecewise polynomials going only approximately

through given data points and fulfilling specific continuity conditions [36]. The smoothing spline is similar to SG in that it estimates a model that contains a smoothed value at each observation. It proved to work faster than MA or SG smoothing and is very handy when there is a lot of missing data which can be interpolated smoothly and 'conservative' by the spline function [37,38]. One variant of this difference penalty smoother [39] has been used for pre-processing of univariate signals from the production records in recovery. The filter is based on the penalized least squares minimization of the difference between a signal and its smoothed version and based on work by Eilers [38]. For instance, Figure 3-3 presents cleanup of the signals related to a permeate tank. Signals from sensors located in this tank experience sudden jolts which can most likely be attributed to the emptying regime of the permeate tank. Namely, when there is no liquid around the sensor, the pH increases drastically (Figure 3-3a). For a similar reason, records from the temperature sensor drift from a relatively steady value to the ambient temperature in the production area (Figure 3-3b). Cleanup of those signals starts by trimming the extreme values (marked in red in Figure 3-3) and treating them as missing values. Next, a first-order difference penalty smoother with a selected weighing factor is applied. The obtained effect is satisfying as these types of signals are expected to be quite smooth and steady-state-like owing to a buffering impact of the tank volume.

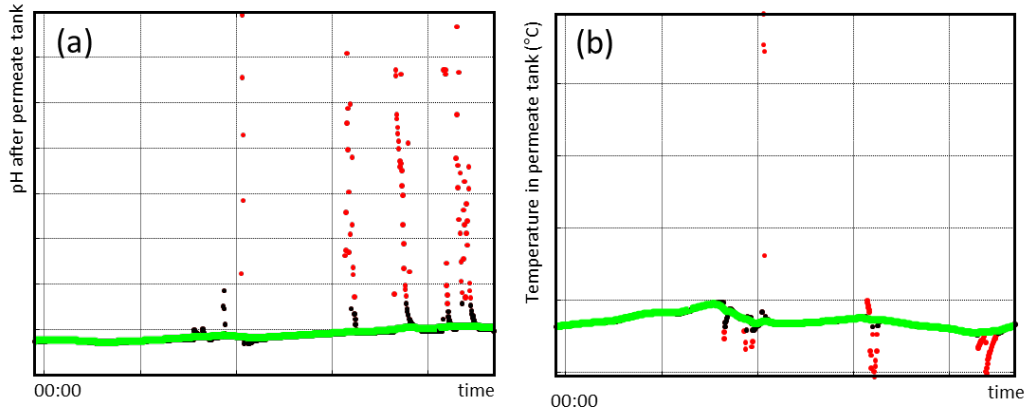


Figure 3-3 – Signal cleanup and smoothing with a difference penalty smoother, (a) pH and (b) temperature. Data marked in red corresponds to the original data which was replaced with missing values. The smoothing spline (green) is fitted to the remaining points (black).

The visual examination of effects of the filters supported by the knowledge about the sensor accuracy, reliability and signal to noise ratio should be used to decide on the appropriate smoothing filter and the correct input parameters for the filters. Signal processing including spectral pre-processing is thoroughly covered in existing literature [36,40].

3.3.2 Stationarity and de-trending

Data from a manufacturing process usually forms a non-stationary time series because in most cases it is governed by the large-scale variation dictated by the set-point values such as production speed. Hence, the statistical parameters - mean and variance - vary in time. Signals need to be stationary with respect to their mean and variance before they can be subjected to lagged correlation analysis. De-trending is a statistical or mathematical operation used as a pre-processing step to prepare time series for analysis by methods that assume stationarity [29]. Generally, it is achieved either through differencing the time series or through the removal of a deterministic trend which is first estimated in a separate regression step [30]. Explicitly, trend removal methods can be assigned to following approaches: 1) differencing (first, second, higher orders); 2) fitting a simple deterministic function of time (least-square fit: straight line, quadratic, exponential, etc.) 3), digital filtering which describes the trend as a filtered version of the original series; 4) piecewise fitting of polynomials (linear, cubic, smoothing spline) [29]. It took many years to study and tease out the statistical implications of these tactics, and still is a challenging aspect of time series analysis [30]. In this project work the first and the last approach have been investigated. Implications of the two trend removing methods are visualized and compared for the same process signal from Figure 3-4 and Figure 3-5.

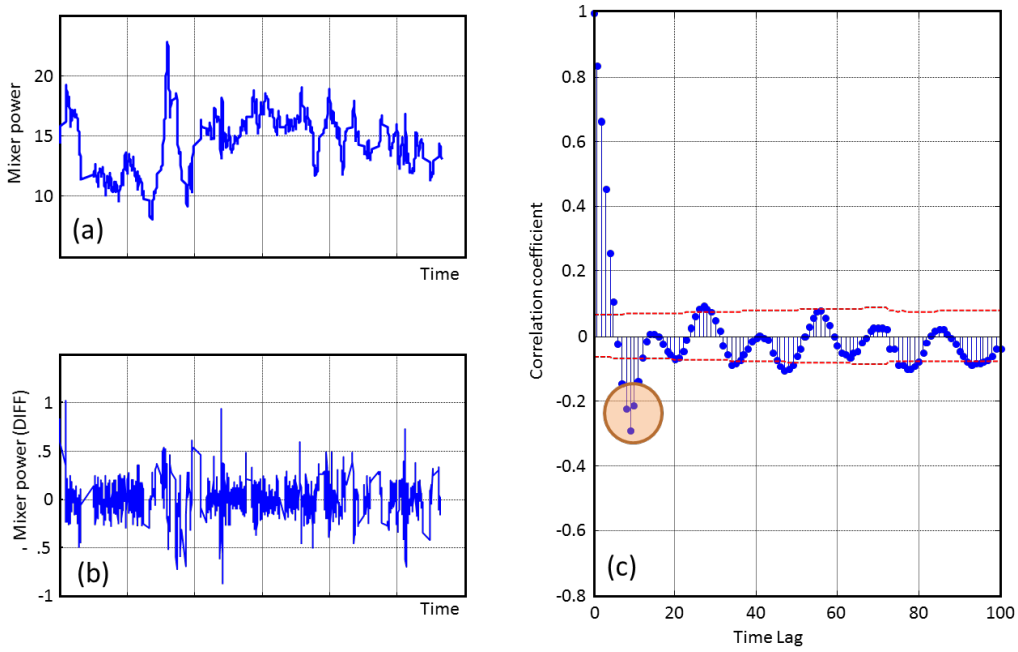


Figure 3-4 - Effect of first differencing on the power consumption records for a granulation mixer; (a) signal smoothed with box-car filter with nine-point window width; (b) differenced signal; (c) ACF of DIFF signal, the red dashed line indicates 99% confidence band.

First the simplest and most common first differencing (DIFF) method was used. Differencing efficiently removes trends and slopes from the investigated signal (Figure 3-4b), however, it was observed that the variance can increase locally, and hence is not constant. ACF of the differenced signal becomes insignificant after lag 5, but it becomes significant again around lag 9 (Figure 3-4c). Even though differencing efficiently removes persistence from ACF, it also induces artifacts which appear to be a consequence of the window size implemented in signal smoothing (Figure 3-4c). This leads to a spurious interpretation of autocorrelogram. The de-trending method which was also easy to apply and offered better results was 'piecewise linear de-trending' (PLD). It is a kind of high-pass filtering where a fitted trend line trails the lowest frequencies, and the residuals resulting from subtracting of that trend line have those low frequencies removed [29]. It comprises fitting straight lines to the data (simple first order polynomials) in sequence, with chosen fixed frequency (Figure 3-5a). Next, the fitted values are subtracted from the original data. The effect is presented in Figure 3-5b. As can be observed in Figure 3-5c, the ACF becomes insignificant after lag 11. However, it turns significant again at later lags which can

indicate that the de-trended series is still not stationary. This can also be caused by previous pre-processing steps or periodicity in the data.

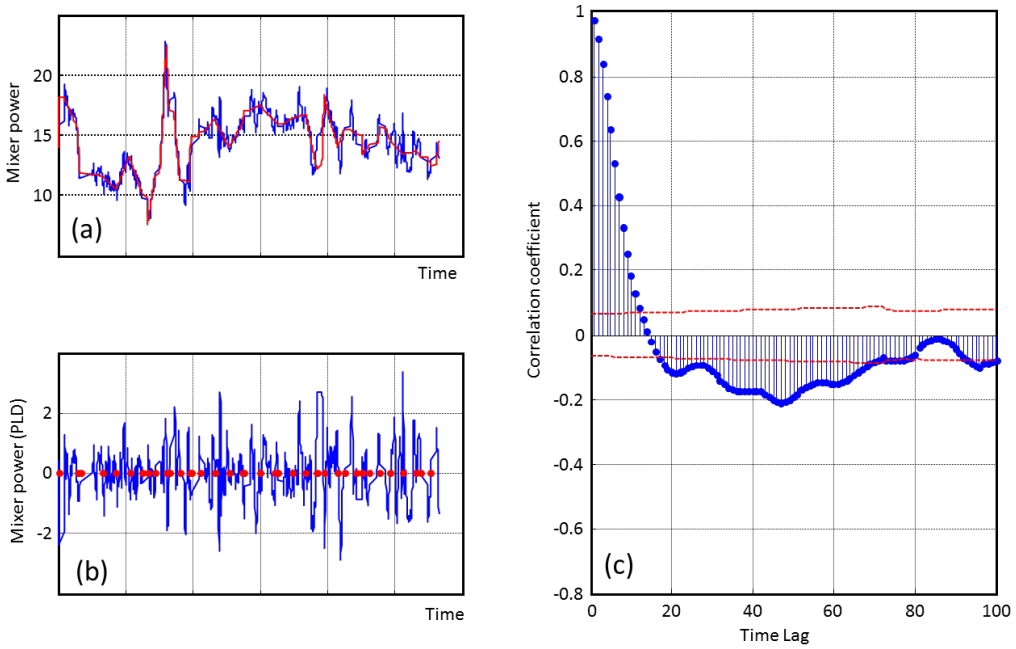


Figure 3-5 - Effect of piecewise linear de-trending on the power consumption records for a granulation mixer; (a) signal smoothed with box-car filter with nine-point window width (blue line), piecewise linear trend fitted every 100 points or if there was a production break; (b) de-trended signal with breakpoints represented by red dots; (c) ACF of PLD signal, red dashed line indicates 99% confidence band.

3.4 Alignment

In any data mining challenge concerning a multi-step production process, the first part of data alignment involves 'Matching'. It has been described in section '2.3 Assembling the data' which covered matching of the information corresponding to one material lot processed in subsequent manufacturing steps. Furthermore, information needs to be aligned within each unit operation which is typically the case in batch production, where it is referred as 'Synchronization'. In the case continuous or semi-continuous processes it is desirable to identify the retention times within one unit operation as well as the delays between the cascade of unit operations which is further referred as 'Lagging'.

3.4.1 Lagging

Delayed reactions are characteristic of many natural physical systems [29] and industrial processes [41,42]. One variable may have a delayed response to another or a delayed response to an impulse which affects both parameters. Also, the reaction of one series to the other series or an external stimulus can be smeared over time, so that a stimulus limited to one observation provokes a response at multiple observations [29]. Serial correlation analysis is an obvious device for studying the relationship between time series. 'Delay' or 'Lag' times of various variables with respect to each other have been reportedly found using cross-correlation [41,42]. However, little or no information is provided regarding the pre-whitening or de-trending procedure applied to the signals before the CCF is used.

3.4.2 Synchronization

One of the principal features of batch processing is its repetitive nature: a certain recipe is consecutively repeated to manufacture batches of a given product. In the chemometric world, much research has been dedicated to correct analysis of the data originating from the batch processes [6,22,43-47]. Synchronization is a first issue that needs to be addressed because rarely are the batches or different, distinct phases over batch runs of the same duration. Differences in batch lengths are observed for instance due to varying effectiveness of catalyst, operational changes, seasonal variations or intrinsic biological variability in microorganisms. Additionally, time points at which the biochemical reactions and physical activities take place may be shifted across batches. Consequently, not only the collected batch trajectories may exhibit various lengths, but also, the key process events may not overlap at the same time in all batches [47]. For this kind of processes, several methods have been proposed to synchronize the trajectories prior to chemometric modeling. Those methods can be assigned to three groups [48]:

1) Compressing/expanding the raw trajectories using linear interpolation either in the batch time dimension or in an indicator variable dimension;
2) Methods based on feature extraction;
3) Methods based on compressing, stretching and translating pieces of the trajectories.

Still, those methods have been developed for the typical batch process. For instance, an indicator variable approach is usually chosen as the simplest and most convenient for industrial applications [13]. In this method, the trajectories are plotted not with respect to time, but with respect to another variable that must be strictly monotonic, has the same start and end values for all batches, and is not too noisy. Thus, an indicator variable can be seen as a pseudo-time. Examples of indicator could be an energy balance, the extent of a reaction or the cumulative amount of reactants added. Next, a constant increment is selected, and one progresses along the indicator variable. Synchronization is performed by retaining the points in the trajectories that have the same values of the indicator variable [22]. Using a 'pseudo-time' instead of the true time has also been applied to some parts of continuous processes [49-51].

3.5 Data reduction

Before or instead of alignment it is possible to limit the number of features within the signal. For example, in the case of process chemometrics, the landmark feature extraction approach tries to capture the relevant information in the evolution of a batch by defining the characteristics of landmarks in process variable trajectories and by recording the values for these features for each batch run [52]. Each trajectory can be split into increments (phases) in which the curve could have different statistical properties. Furthermore, the phases can be identified by certain landmarks that match between the runs (local extremes) and which can be characterized, for instance, by intercepts and slopes [53]. It is not uncommon for a batch process that it runs through different phases (e.g. the lag phase, the exponential growth, stationary phase, and death phase in penicillin fermentation) [54]. Process dynamics and correlations among variables also tend to change with the transitions between the phases. On the other hand, in a well-controlled continuous process, it should be sufficient to approximate the process variables by their means, cumulated values or run length. When signals are less stable it may be informative to add other characteristics that appear relevant such as a standard variation, range or a slope. It is also an option to dissect distinct stages in the continuous process, for instance: a startup, a (quasi-)steady-state and a closing phase (run off), and describe them with separate set of features.

Another option for data reduction is to turn continuous data into binned data, by grouping the events into specific ranges of the continuous variable(s) [55]. It can be

performed in the number of ways depending on the application [56,57]. Similarly to smoothing, data reduction is fraught with information loss. Therefore, it should be preceded with a thorough historical data analysis and used sensibly [58]. In the UF capacity study presented in POSTER II, binning has been used for the reduction of flux profiles. Since these profiles considerably vary in length, it poses a problem for many standard chemometric algorithms to use them as they are without any kind of equalization. Flux profiles had been transformed to bins in the following way: 1) the flux range encountered in historical production runs had been split into ten intervals of equal width (= bins); 2) for each run, the flux entries in each bin had been summed; 3) values in each bin had been normalized by dividing by the total number of entries in that run (= filtration length). Following the binning procedure, the flux in each run was represented by ten fractions corresponding to normalized flux distribution. In the next parts of this thesis, the approach where runs are approximates not simply by their means over the duration of the process but also by other relevant features is called 'Features Extraction' (FE).

3.6 Data validity check

Quality of the data is critical for reliable results of empirical models as these methods relay on data only rather than e.g. chemical or physical insight [12,51]. Assembled datasets should be validated in line with the current process knowledge to establish if they represent the true picture of the process behavior that needs to be explored. A quick preliminary examination of representativeness of extracted process signals can be done for example by 1) estimation of downtime vs. steady-state production time; 2) taking the mean or trimmed mean of the signals per run and plotting it against production time; 3) plotting against other responses that are expected to be correlated. The preliminary examination should involve and facilitate: 1) identification of the missing data; 2) detection of possible mistakes encountered during data acquisition; 3) get an overview of the worst performing runs (downtime); 4) capture sensor failures; 5) spot other abnormalities or errors; 6) check if the information content is sufficient e.g. the observability of investigated fault or quantity to be predicted. If needed, one can return and correct the data acquisition procedures or use the acquired knowledge in the following steps.

The chemometric model building is an iterative process. Therefore, a model can be built already at the early stages on the imperfect data set. In general, examining data

in the projection spaces defined by a small number of latent variables is helpful for understanding the behavior of the process [1]. Such exploratory, preliminary modeling can readily help to identify clusters, outlying runs or abnormal signals. For instance, it would be expected that two pH probes located in the same tank overlap on the loading plots. Frequently such background knowledge of the process is available, and it is anticipated that some parameters will co-vary. One can also figure out that there are several process tags offering similar or the same information and decide to select only some of them or weight them appropriately during the later modeling stages. As an example of data exploration Figure 3-6 presents an overview of the interrelation between different process tags related to a UF operation provided by the Common Components and Specific Weights Analysis (CCSWA) method [59] (see Paragraph 4.4 and Appendix 1 for details).

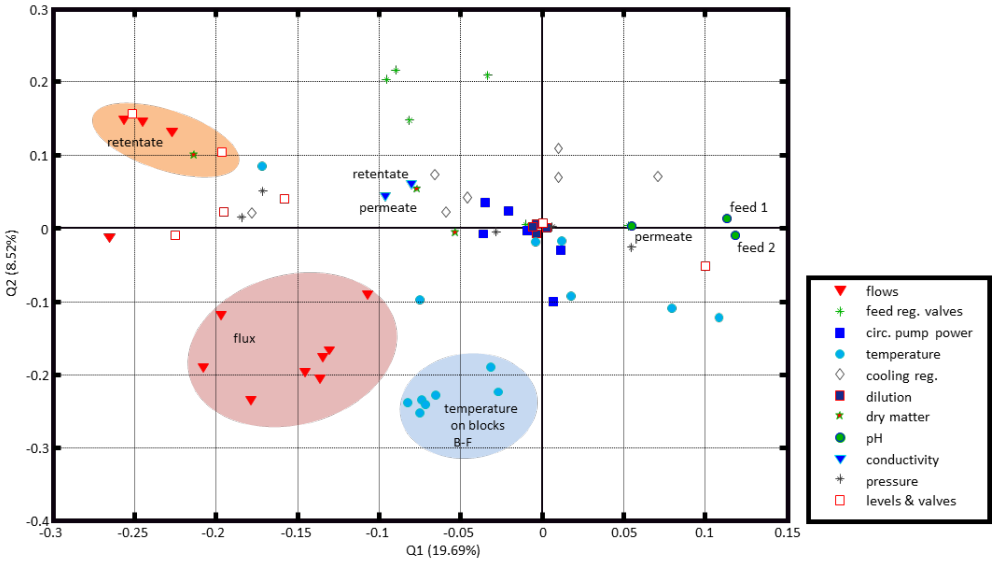


Figure 3-6 - Overview of the association between different process parameters involved in a continuous ultrafiltration operation.

It can be quickly noted that the two pH probes situated in one feed tank do not perfectly overlap (but are close). Mean values recorded by each probe over the same runs are plotted against each other in Figure 3-7a. It appears that the second pH probe, or possibly its position in the feed tank, cannot be trusted as it frequently drifts into unrealistically high values. Another example of the use of simple scatterplot for comparison of supposedly correlated signals is shown in Figure 3-7b which relates the conductivity in the UF retentate and in the UF permeate. Data is

colored according to the production date. The relation between the two measurements looks peculiar. It is possible to distinguish three distinct production periods which vary in the span of retentate conductivity values. The oldest runs are colored in blue. Later, the conductivity of the retentate falls into a lower range (marked with the green-to-orange). This drift can be contributed to some changes in the processing recipe. However, the last group, marked in red, is characterized by a shift in the retentate conductivity to very low values. This was followed by a period with no signal (not shown). As there were no recipe changes that could explain this shift, and the permeate conductivity remained in the previous range, it was decided to set all the values from the retentate sensor to missing for this last period during the analysis.

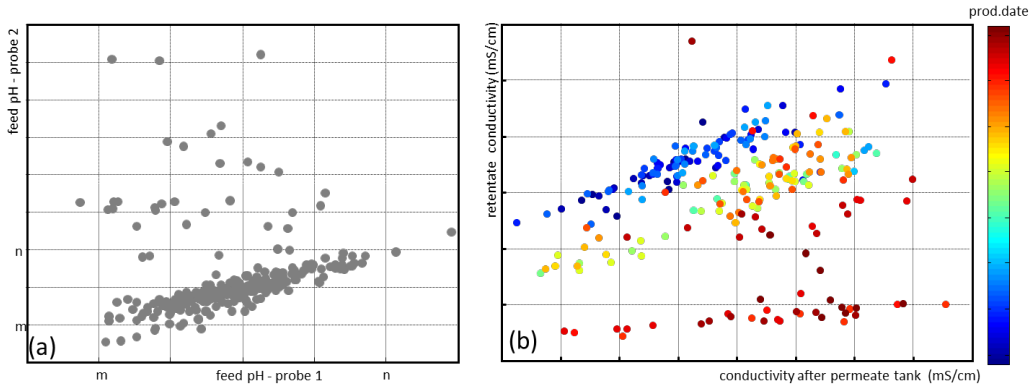


Figure 3-7 – Data validity check: (a) two pH probes situated in one tank; (b) conductivity of permeate vs. conductivity in retentate after the same unit operation; data is colored according to the production dates.

Engineering knowledge about the system under study should always be consulted when evaluating and selecting data. As an example, Figure 3-8 shows the readings for three tags recording UF retentate flow. These three tags form a cluster (together with the corresponding regulation valves) in the plot presented in Figure 3-6. Some process sensors which show very similar information are less reliable than others, and a selection could be made using this preliminary engineering insight. In practice, tag 'YFT03' is the most important. It is used to control the degree of concentration during the UF operation.

Moreover, its readings are logged to the historian with a high frequency. Tag 'YFT04', which is located further downstream, is just a 'back-up' option. This has a

consequence in its logging frequency which is approximately three times lower than for 'YFT03'. Flow meter 'YFT05' also has a high logging frequency, but it experiences a saturation effect which is driven by the requirements of the adjacent instrumentation. Consequently, records from process tag 'YFT03' should be preferred in modeling.

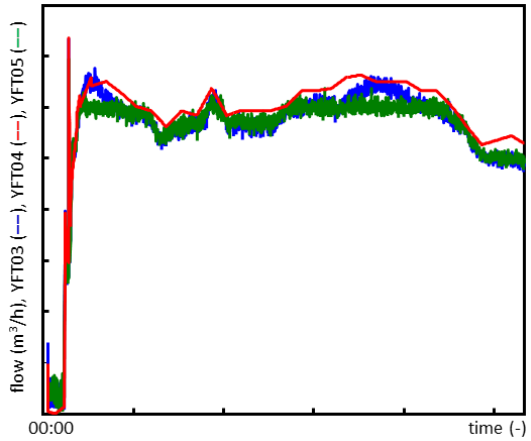


Figure 3-8 – Available process tags recording retentate flow from the UF for one selected run (startup included).

3.7 Utilizing theoretical and process related knowledge

Common sense implies that empirical methods become more powerful when combined with process related knowledge and theoretical or first principle models. Utilizing such knowledge can help determine which variables to include, calculate new variables (transform raw data), decide on the frequency of data sampling and how to weigh the variables prior to model building. Examples that appear in the literature are mass balances, the extent of reaction, and cumulative values. Process specific information (e.g. operator shifts, holding times) can also be included [1]. If calculated parameters are incorporated, care should be taken because measured variables may enter the model several times which increase the weight of this information type in the model and/or monitoring scheme [49]. An example of the use of process knowledge in the UF capacity study (Section 5.1.5 and PAPER III) is the exclusion of all the flow tags and corresponding valves connected via the flow/ratio controller from the regression models for flux prediction. Those parameters show the highest correlation with flux, but this is obviously a causal relation imposed by the two main controllers: the retentate valve controller (which dictates the concentration

degree) and the pressure controller. Flow tags should, therefore, be excluded from the regression analyzes as they would dominate the model and would not lead to any new findings. However, in the exploratory approaches (Section 5.1.5 and PAPER II), flow tags can be kept to have a closer look at the degree of correlation between them and other types of process parameters.

4. Latent variable methods

Latent variables (LV) methods make use of the main characteristic of process databases, namely that even though the it frequently comprises of measurements of a large number of process tags (thousands for the 'factories' in Figure 1-1), the effective dimension of the space in which they vary in a systematic way is significantly smaller (usually between two and ten) [13]. Besides, essential information lies often not in any single process variable but rather in how the variables co-vary [60]. Latent variable methods exploit the above features of process datasets by projecting the high-dimensional data space onto the low-dimensional latent variable space. The latter represents the original data as well as possible, by accounting for the maximum amount of variance. Moreover, LV methods are known to be efficient in separating information from the noise (a kind of signal averaging). They are favorable when a clear understanding of the data is missing, and a large amount of noise is present in the data [60]. Problems of process analysis, optimization and monitoring are thus greatly simplified when working in this low-dimensional space of the LVs [13]. Finally, it is typically much faster to develop a data-driven model than a mechanistic model in a complex engineering task. Consequently, such models (also called 'soft models') enhances the understanding of the fundamental phenomena and processes and often leads to the solution of the problem much faster than traditional first principle modeling approaches. In addition, data-driven models are based on the data measured at the specific processing plant and thus describe the true process reality more closely [2].

This chapter starts with an introduction to the most common chemometric tools which have a global applicability in different, not only chemical, fields of sciences. Next, the specific solutions are outlined which were found particularly suited for the analysis of the large industrial datasets.

4.1 Principal Component Analysis (PCA)

Principal Component Analysis is the key method of Multivariate Data Analysis [61-63], and dimension reduction of the data can be achieved by generating a small number of Principal Components (PC). The PCs are often called underlying (or latent) components, and their magnitudes (concentrations) are called the scores. The principal components are linear combinations of original variables, and the principal component loadings describe the orientation of the PCs with respect to the original variables. The components are determined orthogonal (uncorrelated) and explain as much of the total variance of the original variables as possible up until all variance is explained (Figure 4-1). Normally the first few PCs are used in dimensionality reduction, which is sufficient to cover the highest amount of systematic variation and the most dominant correlation structure among the investigated parameters [10]. The remaining least significant components may simply try to model the unstructured information such as noise.

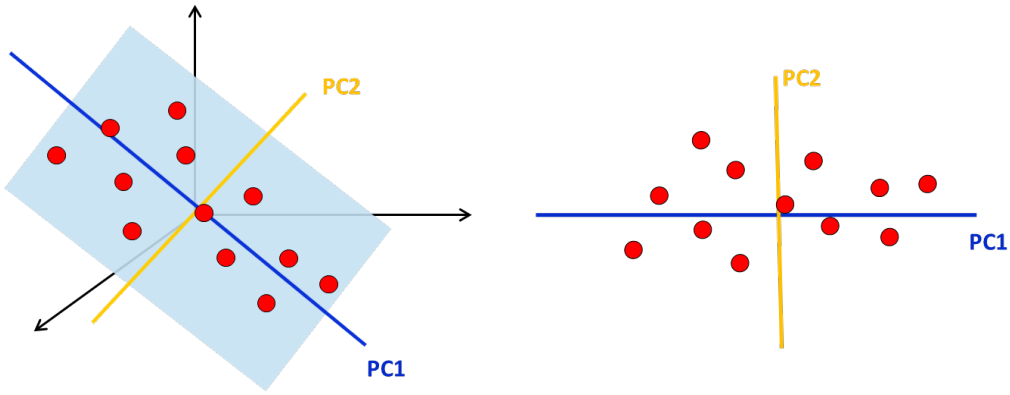


Figure 4-1 - Illustration of the working manner of PCA. The first PC explains the maximum amount of variance in the data set or in the other words spans most common direction in the data. In the same way, a second factor is determined, where the new coordinate is perpendicular to the first one.

PCA holds a strong exploratory and visualization potential. It is e.g. possible to explore subpopulations in a data set by using the scores and loading bi-plots which are two-dimensional windows into the original data set. The loadings indicate which variables are mainly varying among the samples on different PCs and the direction compared to zero. A score plot illustrates the distribution of observations in this plane of the model. The scores provide information on to which extent the variation represented by the loadings are high or low for particular samples.

In summary, PCA decomposes the original data matrix into the multiplication of loading (holding information on variables), score (samples) and residual matrices (Figure 4-2). PCA will show which process parameters (e.g. temperatures, pressures, conductivity, etc.) carry related information, and which of them describe unique variation. The details of PCA algorithm can be found in standard chemometric articles and textbooks [63,64].

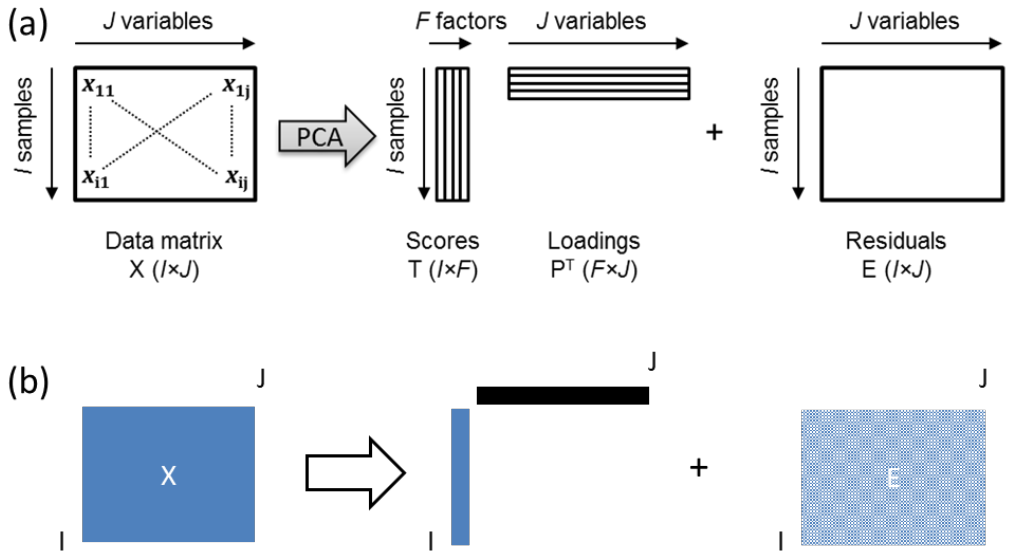


Figure 4-2 - Decomposition of a data matrix by PCA: (a) in matrix algebra notation; (b) pictogram-like notation. PCA defines loadings to filter the noise from the interesting directions and expresses the samples in a new set of variables/scores. Scores form an orthogonal set T, describing the relationship between samples. Loadings form an orthonormal set P, describing the relationship between variables. Variance not explained by any of the principal components, *F*, forms the residual matrix E, which contains noise and redundant information.

PCA has been successfully applied to analyze and monitor continuous processes [65-68]. In this tactic, every sampling time is represented by a row vector of a length equal to the number of process variables and the number of rows is determined by the time-horizon included in modeling. Data from each steady-state production run can also be averaged so that one row corresponds to one process run. As was suggested in Chapter 3, the PCA framework can be helpful already at the early steps of data mining for the evaluation of the validity of the collected data. However, standard PCA can be insufficient or even inadequate for more extensive analysis of

the datasets which consists of several process runs (or campaigns, or batches). This is because it does not take into account the ordered or blocked nature of the data. Explicitly, standard PCA confounds the variation between and within individual data blocks. Different data arrangements and data handling (scaling) prior to the application of the PCA algorithm have been proposed to overcome this issue. This will be discussed in more detail in the following sections of this chapter.

4.2 Partial Least Squares (PLS) regression

Regression problems can be stated as finding the connection between the independent variable $\mathbf{X}$ (size $I \times J$) and the dependent variable, $\boldsymbol{y}$ (size $I \times 1$), or block of variables, $\mathbf{Y}$ (size $I \times Q$). The PLS model can be expressed with a regression vector ($\mathbf{y} = \mathbf{Xb}$) found from a least squares solution. If there are two or more columns in $\mathbf{Y}$, the PLS regression is referred as PLS2 [69]. The acronym PLS can be translated to 'projection to latent structures'. This expansion delineates the idea behind PLS modeling, which as in PCA is taking many variables and projecting them into a lower dimensional latent variable space. The second meaning, 'partial least squares', refers to the way in which the model parameters are calculated. Explicitly, the PLS model fits the covariance between the $\mathbf{X}$ and a corresponding response matrix $\mathbf{Y}$. The response matrix $\mathbf{Y}$ contains the measure of the sample in terms of quality, capacity, concentration or others. In other words, PLS is supervised and can extract latent variables that explain the large variations in the process data $\mathbf{X}$ that is most predictive of the variables in $\mathbf{Y}$. Algorithmic details of PLS can be found e.g. in Wold et al. [70].

4.2.1 Model building and validation

First of all raw data used in a calibration set should be examined visually to detect the most obvious outliers. During the model building phase the most commonly consulted statistics are the Hotelling's T^2 and the Q residuals (Q) [60]. These two tools are used to identify and diagnose outliers during model development as well as afterward when the model is employed. Hotelling's T^2 statistic is a measure of the variation within the PCA or PLS model.

It is the sum of normalized squared scores and a measure for each sample is given by:

$$T_i^2 = \mathbf{t}_i \lambda^{-1} \mathbf{t}_i^{\mathrm{T}} = \mathbf{t}_i (\mathbf{T}_F^{\mathrm{T}} \mathbf{T}_F)^{-1} \mathbf{t}_i^{\mathrm{T}} = \mathbf{x}_i \mathbf{P}_F \lambda^{-1} \mathbf{P}_F^{\mathrm{T}} \mathbf{x}_i^{\mathrm{T}} \qquad (4\text{-}1)$$

where $\mathbf{t}_i$ refers to the i^{th} row of $\mathbf{T}_F$ which is the matrix of F sores vectors from the model and λ, a diagonal matrix containing the eigenvalues (λ_1 to λ_F) corresponding to F LVs retained in the model. T^2 is used to identify extreme points within the LV-space which are generally unwanted as they force LVs to orient in their direction.

The Q residuals are a measure of the amount of variation in each sample not captured by the latent variables used by the model. It is also referred to as the Squared Prediction Error or DModX (Distance to the Model in **X**-space) and defined as sum of squares of each row (sample) of a residual matrix **E**. Hence, for the i^{th} sample of **X**, $\mathbf{x}_i$:

$$Q_i = \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i(\mathbf{I} - \mathbf{P}_F\mathbf{P}_F^{\mathrm{T}})\mathbf{x}_i^{\mathrm{T}} \qquad (4\text{-}2)$$

where $\mathbf{e}_i$ is the i^{th} row of **E**, $\mathbf{P}_F$ contains F loadings retained in the model and **I** is the identity matrix. Q-statistics is used to identify the (moderate) outliers which break the correlation structure described by the model. For both Q and T^2, when a deviation is detected, it is possible to backtrack through the model to identify which variables mostly contribute to the deviating behavior [60].

Once a model is built and saved, it is capable to transform the raw process variables into quantitative measures which can be used for prediction of **Y** values for the new samples. A risk is always present that the identified correlations are caused only by chance and not by true changes in the analyzed parameter. Moreover, a substantial risk of 'over-fitting' exists when using numerous and correlated **X**-variables. This means that the model fits well to this particular data set but has no predictive power when applied to new data. Consequently, it is essential in empirical model building to use some measure of model performance and find the correct model complexity. This can be fulfilled by an appropriate cross-validation (CV) method or, even better, by independent test set validation. The new values for either **X** or **Y** can be tested for their uniformity with previous observations, and can be used to predict new values for **Y** from **X**, respectively. The following measures are the most popular for the evaluation of the model performance:

1) Root mean square error of calibration (RMSEC) is a measure of fit and determines the average deviation of model estimates from actual values:

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I}} \quad (4\text{-}3)$$

2) Root mean square error of cross-validation (RMSECV) is a function of how the model will behave on calibration samples temporarily kept out during model building. It is an estimate of predictive power on the new data:

$$RMSECV = \sqrt{\frac{\sum_{k=1}^{K}\sum_{i=1}^{I_k}(y_i - \hat{y}_i)^2}{I}} \quad (4\text{-}4)$$

3) Root mean square error of prediction (RMSEP) is used to validate the model and is a true measure of predictive power on new data:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I_p}(y_i - \hat{y}_i)^2}{I_p}} \quad (4\text{-}5)$$

4) Squared correlation coefficient (R^2) is the amount **Y** 'explained' in terms of sum of squares:

$$R^2 = \frac{\sum_{i=1}^{I}(\hat{y}_i - \hat{\bar{y}})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I}(\hat{y}_i - \hat{\bar{y}})^2 \sum_{i=1}^{I}(y_i - \bar{y})^2}} \quad (4\text{-}6)$$

where:

k : subset numer used in cross-validation

I: number of calibration samples

I_k: number of samples in a CV subset

I_p: number of samples in a test set

y_i : measured value for sample i

$\bar{y}$: mean measured value

$\hat{y}_i$: predicted value for sample i

$\hat{\bar{y}}$: mean predicted value

5) Cross-validated R^2 (R^2 (CV)) - the amount of **Y** 'predicted'.

4.3 Centering and scaling

Another important step before data is subjected to multivariate analysis is pre-processing of the raw data. Centering and scaling are the two most commonly applied techniques. Centering is performed to remove the offsets, and it corresponds to re-positioning of the coordinate system so that the average point now is in the origin. Without centering the first factor in the model would be used to explain the distance from zero to the center of the data cloud (Figure 4-3a). Centering is usually performed across the first mode (rows = samples). This involves subtraction of the column means from the elements in the corresponding columns of an $I \times J$ matrix $\mathbf{X}$ and produces the matrix of deviations from column means or the column-centered matrix $\mathbf{X_C}$:

$$\mathbf{X_C} = \mathbf{X} - \mathbf{1m^T} \qquad (4\text{-}7)$$

Where $\mathbf{1}$ is $I \times 1$ and $\mathbf{m}$ ($J \times 1$) is a vector containing means of column j as its jth element.

Variance scaling is used when variables are in different units or different magnitudes. It is used because variables with little variation would not be modeled to any significant degree whereas variables with the highest variance would prevail in the model solution. Centering does not remove the scale differences between variables; it just centers the variation around zero. Since the difference in scales between variables is arbitrary, it is convenient to scale the data so that each variable has the same preliminary standard deviation and simultaneously removes different measurement units [71]. This can be obtained by scaling the centered data within the second (columns = variables) mode, where every column of the centered matrix $\mathbf{X_C}$ is multiplied by a specific number:

$$\mathbf{X}_{CS} = \mathbf{W}\mathbf{X}_C \qquad (4\text{-}8a)$$

where $\mathbf{W}$ is a $J \times J$ diagonal matrix with the scaling parameter for the jth column on its jth diagonal. The weight of a variable is often selected to be an inverse of the standard deviation (s_j) of that column:

$$w_{ij} = \frac{1}{s_j} \qquad (4\text{-}8b)$$

An alternative to variance scaling could be e.g. 'range scaling' or a less severe version like Pareto scaling where the square root of s_j is used. The combination of centering

across the first mode and variance scaling within the second mode is often referred as 'auto-scaling' (Figure 4-3c).

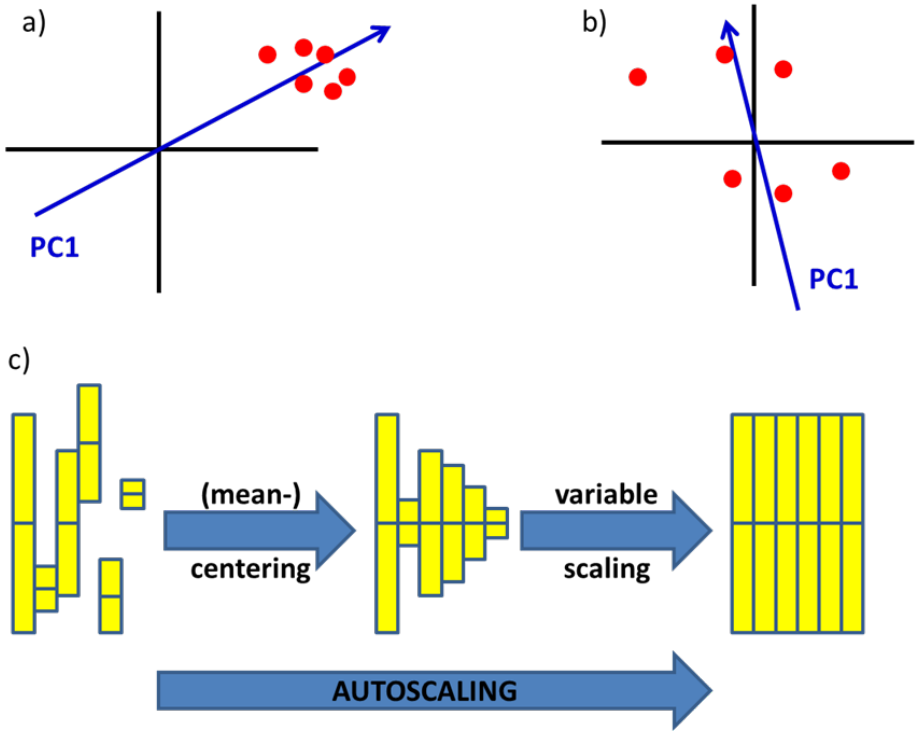


Figure 4-3 – Effect of mean-centering on the determination of the first PC: (a) before mean-centring, (b) after mean-centring. After auto-scaling (c) all variables have equal 'length' and mean value zero.

When data consists of variable subsets of a significantly different size, or of a number of conceptually relevant variable blocks, it can be beneficial to scale each block separately to ensure that all the different blocks are allowed to contribute to the model. This is a crucial step in multiblock modeling (see Paragraph 4.4). If standard auto-scaling is performed blocks with fewer variables will have less influence in a multiblock model and the opposite happens for the blocks with many variables. The solution is individual block weighting, or 'block-scaling', where it is possible to downscale large blocks and upscale small blocks. Variables within each block can be mean-centered or auto-scaled, after which different blocks can have their block-variance normalized to the same value.

4.4 Dataset arrangement and unfolding

The key difference between alternative techniques explored during this Ph.D. project lies in the form in which data is arranged (Figure 4-4). This preparation step involves very distinct operations that all lead to the presentation of the data in the form of a two-dimensional (2-D) matrix. This is done because most of the chemometric methods (all based on the two fundamental algorithms PCA and PLS) can work only if the data is rearranged in this manner. In the standard setting of process chemometrics, columns of the data table corresponds to process parameters/tags (J) and rows are the process timestamp (N_i). With this in mind, the simplest way to decompose the data table is to analyze one run at the time (Figure 4-4, case A). However, in this approach, it is not possible to directly compare the model outcomes over different runs/batches as scores for each decomposition - due to rotational freedom – will result in an individual set of loadings. Therefore, scores corresponding to different timestamps cannot be compared between the runs. Nonetheless, if there is a need to explore the main sources of systematic variation in one problematic run and to analyze which timestamps appear as outliers, then approach A might be valuable. A direct comparison of scores from different PCA solutions can be possibly done after the loading matrices are rotated to a similar and simpler structure using e.g. VARIMAX rotation [72]. This rotation does not change the sum of squared residuals, under the condition that counter-rotating of the scores compensates the rotation. The obtained components can be orthogonal (uncorrelated) or oblique (correlated), depending on the rotation technique used [73].

Alternatively, the common association between the investigated parameters can be found by CCSWA method [59] (Figure 4-4, case B). In this approach, data tables are expressed in terms of cross-products ($\mathbf{W_i} = \mathbf{X_i X_i^T}$). The association matrices, $\mathbf{W_i}$, reflect similarities between the process parameters for each run. The CCSWA looks for a common loading (q) which explains as much variance as possible for all association matrices, via weighting (saliences, λ). This is done in the iterative way by eigenvalue decomposition of the weighted-average association matrix. The algorithmic details of this method are given in Appendix 1. In the end, CCSWA offers the graphical display of the mean configuration of the investigated parameters on the basis of derived components and of the data sets on the basis of saliences (as used in POSTER II). Saliences reflect how the $\mathbf{W_i}$ configurations are merged, hence, to which extend a specific data set agrees with the latent variable shape described by

the particular component. The main advantage of CCSWA is that datasets do not need to have an equal number of rows (time points). However, as a consequence of the formation of cross products, the time dimension is lost in CCSWA. Several alternative techniques exist which allow us to concatenate information from different process runs into 2-D matrix suitable for bilinear modeling (Figure 4-4, cases C-F).

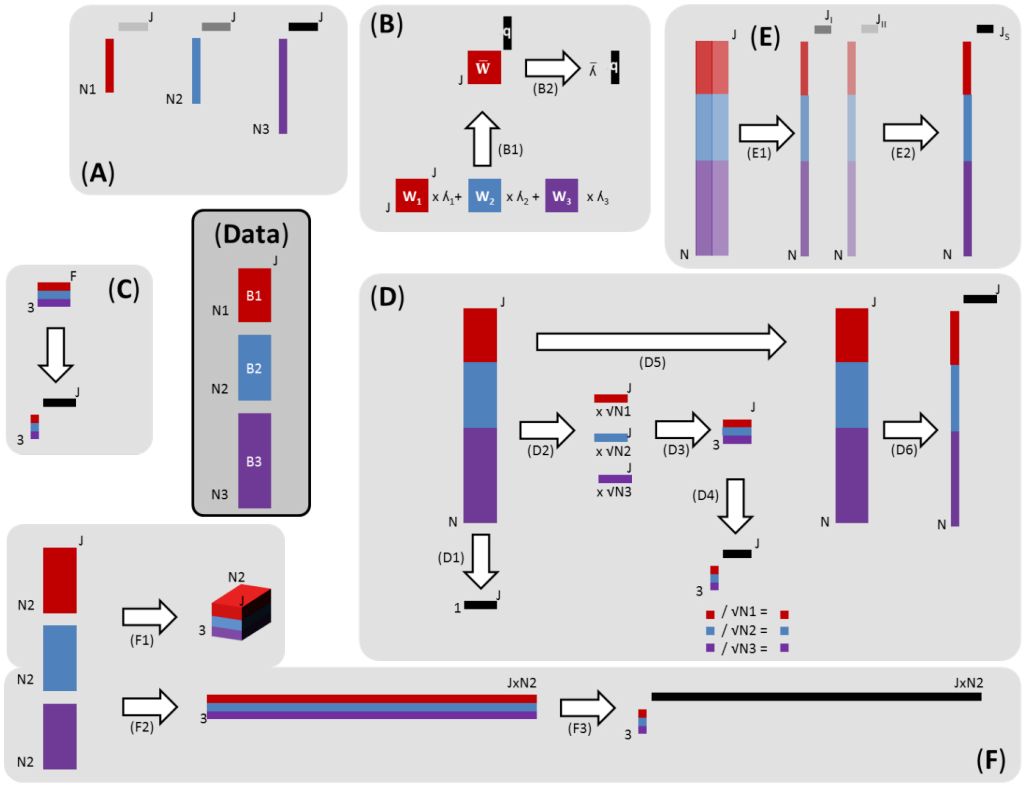


Figure 4-4 - Different types of data arrangement and modeling: A) Separate decomposition of individual batches, B) Common Components and Specific Weights Analysis, C) Features extraction (number of features, $F = \sum_{j=1}^{J} f_j$, where f_j is the number of features selected to express a process parameter), D) Variable-wise unfolding and multilevel modeling, E) Variable-wise unfolding and multiblock modeling, F) Batch-wise unfolding and modeling.

The conceptually simplest is to extract process features as it is described in Paragraph 3.5. Selected features form columns and one run is reduced to one row in the new formed matrix (case C). Alternatively, several datasets that have a common variable mode can be stacked below each other (case D and E). If two modes are common between the datasets (case F), then it is possible to stack the dataset on the top of each

other, as it is shown in step F1. With this type of data organization, the sequential nature of the data, hence how it is generated plus organized, can be accounted for using the modeling techniques referred to as multiway methods [60]. The latter arrangement can also be accomplished after some kind of data equalization (discussed in section 3.4.2 'Synchronization').

In process chemometrics, the three-way data structures are typical used for batch processes. As a consequence, it is custom to differentiate between the two practical ways in which such data cube can be unfolded to 2-D form, namely [47]: 1) 'Variable-wise unfolding' (VWU; case D and E), also called 'Observation-wise unfolding'; and 2) 'Batch-wise unfolding' (BWU; case F).

Furthermore, it can be beneficial to utilize information on the conceptually meaningful groupings recognizable either in the row or variable dimension of augmented dataset. When rows are nested within groups (organized hierarchically) and share the same variable mode (case D), Multilevel Simultaneous Component Analysis (MSCA) [74] is a suitable LV method. In this approach, patterns belonging to different levels in the data hierarchy are modeled separately. If variables can be broken into meaningful blocks (case E), for instance with each block corresponding to a processing unit or a section of a unit, then multiblock modeling techniques should be used [75-77]. Both multilevel and multiblock models have improved interpretability over the conventional LV approaches [16,22,74,78-80].

Figure 4-4 sketches the major steps involved in the above techniques and they all end in a PCA-based decomposition. Naturally, the PLS versions for all data arrangements (except the CCSWA) are also possible if a reference or target y-value is available. It is important to understand the advantages and disadvantages of each method and which of these approaches perform successfully on what types of processes data. This is explicitly discussed in the Paragraph 4.6.

In the framework of this Ph.D., it is desired to use a large number of data collected over several years of production, and hence efficient methods capable of comparing numerous different runs are needed. Data organizations and corresponding model techniques that fulfill these requirements are C, D, E and F (Figure 4-4).

4.5 Variable selection

Variable selection is about limiting the number of variables to the most informative ones. The use of apriori knowledge, e.g. choice of process tags determined by engineering intuition, is the most fundamental and influential variable selection strategy [81]. However, it might also remove relevant or important information in an exploratory multivariate modeling due to a predisposition in the understanding of the production system. Variable selection can be driven by the following motivations: 1) an improvement of the model predictions (by removal of many irrelevant, noisy or unreliable variables); 2) a better interpretation for instance by removing all variables that do not contribute significantly to the model (reduced model complexity); 3) an improvement of statistical properties of the model; 4) minimise the risk of over-fitting 5) decrease computational time; and for later use 6) to reduce the costs in measurements. The variables used in multivariate data analysis should be considered as a whole and not by looking at one variable at the time as frequently a variable is useful in prediction only in combination with other variables included in the same set. It can also be decided to remove variables that are strongly correlated. Even if it does not result in improved predictions, it offers a simpler model. Some variable selection methods are based on an assessment of minor differences in quality of the model and even in assessing the significance of statistics calculated from the model parameters (such as predictive performance). Therefore, it is recommended to remove even minor outliers prior to variable selection. In this way, it is assured that the variables were selected not only due to extreme behavior of some samples. Outliers should be investigated in both dependent (**X**) and independent variables (**Y**). Sometimes a sample may appear as an outlier in the original dataset but not in the dataset limited to selected variables. Therefore, the outlying samples can be reintroduced to the model, and outlier selection might have to be performed iteratively over the selecting process. The variable selection is thus frequently an iterative procedure, where the analyst is working his or her way towards a good solution [81].

The most popular methods used in variable selection process are:

1) Using model parameters and diagnostics; for instance, variables which have low loadings or low regression coefficients could be removed.

2) Variable Importance in Projection (VIP) is aimed at finding variables that are necessary not only for prediction but also for describing **X** [82].
3) Selectivity ratio (SR) - provides a simple numerical assessment of the usefulness of each variable in a regression model by calculating the ratio between explained and residual variance of the spectral variables on the target-projected component [83].
4) Genetic algorithms
5) Classical statistical approaches
6) Interval Partial Least Squares Regression (iPLS), which calculates separate models for windows of ordered variables (e.g. as ordered wavelengths in spectroscopy) to find one or a few intervals which offer better predictions than then all variables [84].

Note that only engineering insight, 1) and 5) can be used in an unsupervised method like PCA, while for supervised methods many more options are available. In general it holds that variable selection for PCA is more subjective/idiosyncratic while – accompanied by the correct validation methods – PLS plus variable selection is more objective/impartial. In PAPER I, iPLS has been used to discard parts of the spectra which were judged irrelevant for prediction of active enzyme protein. VIP and SR have been used in connection to UF case study (Paragraph 5.1.5) to determine process parameters which showed the strongest correlation to the flux decline.

4.6 Modeling of batch vs. continuous processes

One important question in deciding what type of modeling should be applied is whether the system under study is static or dynamic. Batch processes are by definition dynamic processes [11]. Usually, data from a batch process consists of J process variables measured at N_i points in time along a batch (i). True batch processes can be considered replicates of each other [78], but usually, a synchronization step (Figure 4-4 F) is required, after which a multiple-batch data can be arranged into a three-way tensor $\underline{\mathbf{X}}$ $(I \times J \times N_i)$. Variable trajectories measured along the duration of a batch are non-linear with respect to time, and they form a multivariate time series of a dynamic nature [45]. BWU is the most commonly applied method to unfold a three-dimensional data matrix (Figure 4-4, step F2). It keeps the dimension in the batch direction and merges the variable and time dimensions. Each row of the unfolded matrix, $\mathbf{X}$ $(I \times JN_i)$ contains all data within that

batch and each sample of process measurements at different sampling interval is considered a new variable [43,85].

In a continuous process the $\mathbf{X}_i$ matrix corresponding to one run consists of observations on J variables collected at successive time intervals (N_i). If several runs are available then the data can be easily augmented as in Figure 4-4 D. In the unfolding contex this operation is refered to as variable-wise unfolding (data matrices are stacked below each other, $IN_i \times J$). The covariance structure between the variables in a steady-state continuous process should remain stable over time, therefore, this type of data is frequenty analyzed just as averages of the process tags over the run. This is a special case of C shown in Figure 4-4 where $F = J$ and $\mathbf{X}$ has size $I \times J$. Variable-wise models incorporate only the variances and instantaneous cross-covariances of the variables [46]. Thus, this modeling strategy is only valid when the correlation structure of a process is more or less constant. The BWU modeling approach has been previously applied to continuous processes but only to their specific parts, such as grade transitions, startups and restarts [49-51]. These sequences all share the same three common stages: the initial conditions, the transition and the final steady-state.

It is custom to differentiate between the variable- and trajectory- scaling and centering depending on the unfolding strategy [46,47]. If data matrices are stacked below each other like in variable-wise unfolding ($IN_i \times J$) the variance scaling is called 'variable scaling' and centering the 'variable centering'. If unfolding is performed batch-wise ($I \times JN_i$), Figure 4-4, case F, step 2, then the variance scaling is referred as 'trajectory scaling' and centering as 'trajectory centering'.

For data from a typical batch process, 'variable scaling + centering' does not remove the average trace from the data. Thus, if the investigated variables are expected to follow a particular trajectory, then 'trajectory scaling + centering' ought to be applied to focus only on the deviation from the average trajectory and not from the global mean of that variable during production. At the same time, 'trajectory scaling and centering' can eliminate 'non-linearity' in the system. Simple transformations of the data such as logarithms may also help with non-linearity [1].

Covariance and partial covariance maps applied to the batch-wise unfolded data are helpful tools for investigation of the process dynamics as they give a picture of the

covariance among process variables over time [86]. A dynamic partial covariance map can be useful to choose a parsimonious model. It visualizes the dynamic relationships between variables without taking into account the direct relationships. This map should be used when the objective is to predict the current value of a variable from previous measurements of the process, for example, to make the one step ahead predictions. A theoretical discussion on the capability of different modeling/unfolding methods to capture the process dynamics based on the structure of the covariance matrices and the use of covariance maps is discussed by Camacho et al. [46,86].

The intermediate methods between BWU and VWU approaches are also available and called the (batch-) dynamic unfolding [46,87]. This type of data arrangements is equivalent to the VWU with lagged measurements added as extra variables. If all possible lagged measurements are added, the resulting matrix is the same as after BWU. Therefore, batch dynamic PCA and PLS models [87,88] can be seen as a generalization of the traditional unfolding procedures.

Other strategies have been proposed for multivariate statistical process control and monitoring including 'local model', 'evolving model', 'adaptive models' either as single models for the entire process, or separate models for distinct phases in the process [6]. However, for mining of complex data sets, as the ones obtained in semi-continuous processes, multiset modeling strategies are of more interest. The generic problem in the multiset analysis is to find underlying relationships between several datasets [89]. Three classes of multiset modeling approaches were of particular interest in the framework of this thesis work. The first of them are the BWU-PCA and PLS (Figure 4-4 F), which can be perceived as multiset methods that deal with datasets that have two modes in common. The second class of problems is encountered when only the variable mode is common between the dataset. It will be addressed with Multilevel Simultaneous Component Analysis with invariant Pattern (MSCA-P) [74]; the algorithm is outlined in PAPER II. The third class of multiset techniques is designed to address situations when objects are in common, but the variables measured on these objects are different or can be split to distinct groups [77]. For clarity, this case is further referred to as a multiblock (MB) problem, although in the literature, multiblock and multiset is generally used interchangeably.

Discussion on available algorithms and motivation for multiblock analysis is provided in PAPER III.

4.6.1 Data-driven methods for process monitoring

In a manufacturing plant, we want to use processes which are stable and repeatable to meet customers' demands (where the customer is often a downstream operation) in terms of product quality as well as to closely follow our production schedule. The process is said to be 'in statistical control' when it is affected only by an unavoidable random ('common cause') variation which is an integral part of the (stable) process and unit operations. This means that it is predictable from statistical moments, and therefore the process outcome in the near future can be foreseen [60]. Typical process monitoring strategies in LV space rely on PCA, PLS or more advanced extensions of these two methods [60]. The choice of the reference set to define common cause variation and its quality is crucial for a successful application [20]. The real goal of the monitoring is to discriminate between normal and abnormal operating conditions. Therefore, only data obtained under Normal Operating Conditions (NOC) should be utilized, and production episodes which contain variations due to special events (Abnormal Operation Conditions, AOC) should be excluded in the model building stage. E.g. samples from Designed Experiments span the extremes of the variation and should not be used for that type of application [60]. If a model is expected to be robust e.g. towards different product variants or throughputs, then it is important to incorporate this range of variation where good predictions are expected. It needs to be remembered that when a data-driven model is used for prediction, extrapolation outside the validated range is in principal unsafe unless proven differently [60].

4.6.2 Multivariate Process Monitoring

A model relating **X** and **Y** is constructed using the available historical or especially collected data. LV control charts can be built to monitor the predictors, taking into account their impact on the response variables. This approach means that the efficiency of the process can be supervised, even when the product quality measures (**Y**) are not available [13]. Furthermore, it is anticipated that monitoring of the process data – via compressed information in the LV space - provides more information on the state of the process than following just (end-)product quality data (**Y**). This is because if any unusual or abnormal event takes place, this will leave a fingerprint in

the process data. Consequently, it is easier to diagnose the source of the problem as when directly dealing with the single process variables [90].

The most common set of diagnostics used in multivariate control charts are Hotelling's T^2 and the Q-statistic as discussed under 4.2.1 [60]. Corresponding control limits can be derived and used to detect out-of-control situations. In particular, Q-statistic serves to detect process drifts and estimate if the model is still valid. Various heuristic rules have been proposed to signal the onset of process faults [91]. Since these charts are only able to pinpoint abnormalities in the process, the corresponding contribution plots are used to identify which process tag(s) is(are) responsible for the rise in Q and/or T^2-values and to direct to the part of the process which was affected by a particular fault. Illustration of the methods, algorithms and details on the estimation of the control limits are thoroughly described by Kourti [13,66], Ferrer [92], Montgomery [93], Westerhuis [94] and Bersimis [95].

Readers interested in knowing more regarding on-line implementations of multivariate statistics and indispensable model maintenance are redirected to the industrial experience of other practitioners [12,15,96].

5. Industrial cases

Enzymes are easily biodegradable proteins which occur in all living organisms. Their role is to catalyze biochemical reactions. From and industrial technology and consumer product perspective, enzyme technology can replace conventional chemicals to improve resource efficiency and reduce environmental impact. In 1952 Novo A/S (later Novozymes A/S) introduced Thermozyme®, the world's first enzyme produced by a fermentation process which unlocked the option to manufacture enzymes on a large-scale [97]. After this first step, many different enzymes have been developed, and today Novozymes A/S is a global leader in production of these biotechnological aids. Major part of the business is development, production, and distribution of enzymes. Novozymes A/S offers over seven hundred products in one hundred thirty countries, applicable to more than thirty different industries. The range of application areas is wide, including for instance food and beverages, paper and pulp, plus textile and household care. Out of the sixteen global sites, five large-scale production facilities produce the majority of Novozymes' biotechnological solutions. Data and processes investigated in this Ph.D. project originate from the production facilities located in Kalundborg, Denmark. Nonetheless, it is expected that the learnings are also valid for other production sites.

The initial step in the production of industrial enzymes is cultivation which involves aerobic submerged fermentation during which enzymes are secreted from cells (Figure 1-1). Enzymes are usually commercially produced using bacteria, yeast or fungi [98]. Nowadays, the production of enzymes starts from a vial of dried or frozen microorganisms that have been selected or genetically modified to yield large amounts of enzymes [99]. After several cultivation steps, the content is transferred to the large bioreactor which operates in fed-batch mode. During the cultivation, numerous operational parameters such as feed rate, oxygen consumption, temperature, and pH are monitored or controlled to secure optimal production conditions. When the main cultivation is completed, the culture broth is cooled down. Next, enzymes have to be separated from the biomass (i.e. cells, nutrients, by-

products), purified and concentrated in the subsequent factory step, called 'recovery'. Finally, the liquid enzyme product is formulated according to downstream processing; the enzyme is either sold as a stabilized liquid product or processed into granulate in a third factory.

Cultivation as a biologically based system is the most complex to control and adjust [100,101]. Recovery and granulation are governed by physicochemical events and can be described easier using modeling. Chronologically, potential fields for improvement have been investigated in this thesis work starting from the granulation process. In the beginning, it was important to comprehend the final product characteristics to understand relevant quality parameters that could be affected along the way. Quality of the final product at Novozymes A/S can be expressed in terms of 1) enzyme activity, 2) stability (deterioration of activity with time), 3) particle size distribution, 4) the amount of 'dust' (small fragments with enzyme activity, which is not only an undesired economic loss but also poses a serious health aspect if inhaled) and 5) color.

The direction of Process Analytical Technology (PAT) research in bioindustry is driven by increased competitiveness and pressure for improved efficiency in production and process development [100]. Thus, the aim of PAT is to support innovation, efficient manufacturing, and quality assurance [102]. Process Analytical Chemistry techniques, such as different types of vibrational spectroscopy, have been playing an important role at Novozymes A/S for over 20 years [103]. These solutions have been used routinely in Research and Development (R&D) and Quality Control (QC) laboratories as well in the laboratories dedicated to production where they offer enhanced process understanding or superior process monitoring. From 2010, the strong pull for PAT from production initiated investigations of the potential of NIR instruments for on-line monitoring of downstream processes involved in enzyme recovery and granulation. At the same time, the role and value of the large volumes of the already collected historical production data in process optimization have been questioned, and the process chemometrics toolbox was proposed as a solution to answer this demand. Consequently, PAT has been identified as a strategic initiative in the corporate strategy at Novozymes A/S. Development of the PAT projects is expected to enhance focus on science in the bulk production, increase process

understanding, reduce variation and minimize the dependency on the time-consuming QC analyzes.

5.1 Recovery of enzyme

When the cultivation is completed the mixture of cells, nutrients and enzymes is subjected to downstream processing, traditionally called 'recovery', where enzymes are separated from the broth. A schematic flow diagram of the process is shown in Figure 5-1. During the purification, cell debris is removed by flocculation and centrifugation. In the flocculation step, the culture broth is first diluted with water to reach constant conductivity required for optimal flocculation conditions. It also ensures an adequately low solids load during the subsequent separation steps. Next a salt, calcium chloride, is added to improve flocculation by neutralizing the negative charges on the cells, stabilizing the enzyme and preventing it from binding to the biomass. Afterward, a polyaluminium chloride is added. The aluminum source serves to aggregate the particles into larger flocs, which makes the flocculation stronger and also helps to remove color from the process. Several polyaluminium chlorides can be used which vary in chemical composition (e.g. basicity and the aluminum content). After the separation step, enzymes are further purified by kieselguhr filtration and then concentrated in two subsequent ultrafiltration steps.

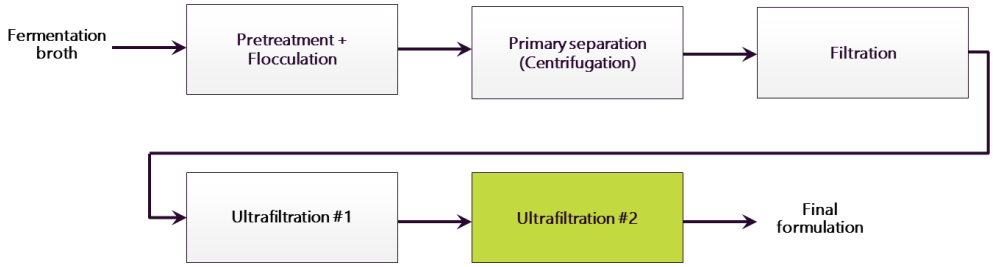


Figure 5-1 - Schematic flow of the enzyme purification process with the key unit operation of the subsequent studies indicated in green. There is usually more than one unit operation involved in each step.

Optimal performance of the recovery process depends on many parameters. Since it is further downstream than the cultivation the final product specifications such as enzyme activity, color, and turbidity naturally become more and more important due to the concentration enhancement. In particular, as the strength of the intermediary product increases, the yield balances over different separation steps become increasingly critical from an economic perspective. Consequently, a good on-line

surveillance of the recovery process is recommended to assure an economic and competitive processing of enzymes as well as to assist in process optimization [103]. The recovery studies described in this thesis are focused mainly on the second ultrafiltration step presented in Figure 5-1. Some upstream information has also been incorporated in part of the studies.

5.1.1 Ultrafiltration system

Ultrafiltration is a technique that gained popularity in different areas owing to its advantage in the separation of molecules without phase or heat transitions [104]. It has been applied in the dairy industry, bioprocessing, waste water treatment, and the paper and pulp industry for a number of years [105,106]. Ultrafiltration in downstream processing of enzymes has two main aims: 1) to separate water (concentration); 2) to filter out impurities (purification, e.g. small molecules: mono- and di-saccharides, salts, amino acids, organics, inorganic acids or sodium hydroxide).

The ultrafiltration unit which is in the center of the studies presented in this thesis is a plate and frame system [107]. Its smallest working element is a membrane. Membranes retain enzyme molecules (based on their size and shape) in the retentate while allowing for the permeation of water and small molecules. In the case of an enzyme recovery process, the retentate is a high-value product that contains enzymes, whereas permeate is a waste product that contains water and impurities. Permeate can however be reused to replace dilution water in upstream processes.

Membranes are polymer sheets (Figure 5-2a) fitted in pairs between supporting hard plastic plates with spacer channels (Figure 5-2b). The pores of the membranes are very small and therefore the ultrafiltration is driven by pressure. The feed is pumped between the surface of the paired membranes, 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 plate spacers, where it is led away through a permeate tube (Figure 5-2c). One membrane module consists of hundreds of membrane sheets and supporting structures (Figure 5-2c). In one module, the direction of the feed flow can be changed. This change is imposed by separating plates. Several modules working in parallel form a recirculation loop, also called a 'block' (Figure 5-2d).

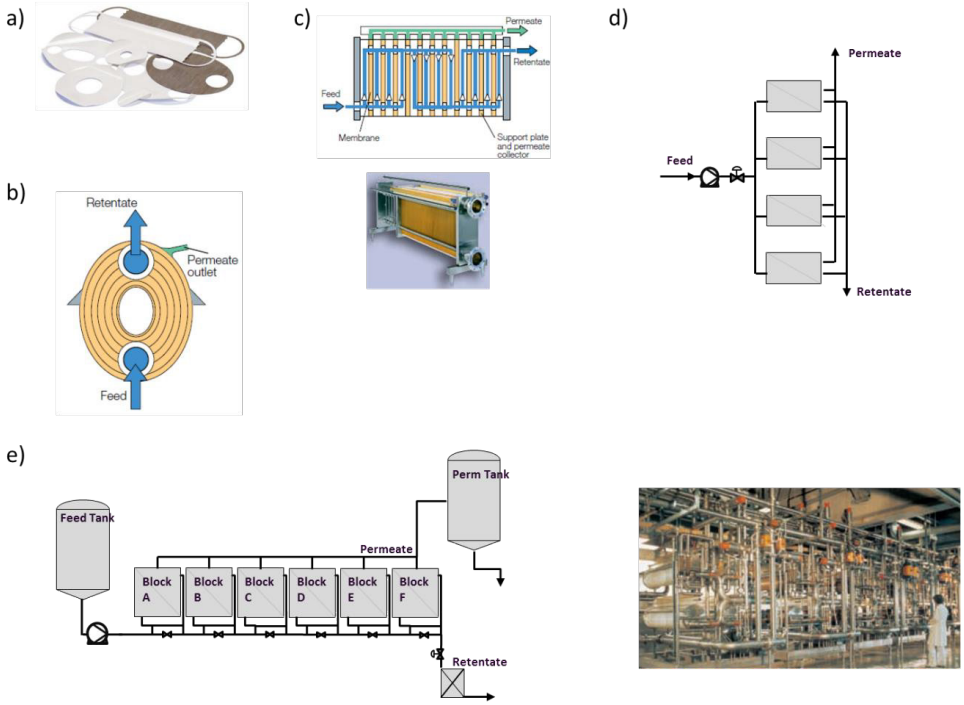


Figure 5-2 - Construction of a plate and frame system for ultrafiltration (see text for details): a) membranes; b) membranes + supporting structures, c) UF module d) UF block/loop e) multistage recirculation plant = UF unit. Photos and schemes taken from [108]. The operation historian collects information on the block level.

Each block of membranes has a centrifugal pump (booster pump) and accompanying throttling valve to provide pressure and ensure an adequate cross-flow velocity of the feed over the membrane. This cross-flow 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. Other key components external to the loops are a feed tank followed by the feed pump, a permeate tank, pipelines and a heat exchanger on the retentate stream. There is a number of flow transmitters installed to monitor and control the throughput.

5.1.2 Operating the UF system

Operating a continuous multi-stage membrane system with several recirculation loops can be quite complex [107]. As is described in Chapter 2, selection of a batch or another increment of the production corresponding to one entity in our data mining

is not a trivial task. In this section, a more detailed explanation of the potential implications of different operation regimes to the structure of the data is delineated. In the UF studies it has been decided to distinguish between the startup phase and a ('quasi'-) steady-state phase during filtration. The startup is initiated when feed enters the plant and the UF outlet is simultaneously closed to allow for a concentration build-up. The startup finishes when the dry matter set-point is reached in the retentate stream which, from this point in time, is redirected further downstream (Figure 5-3).

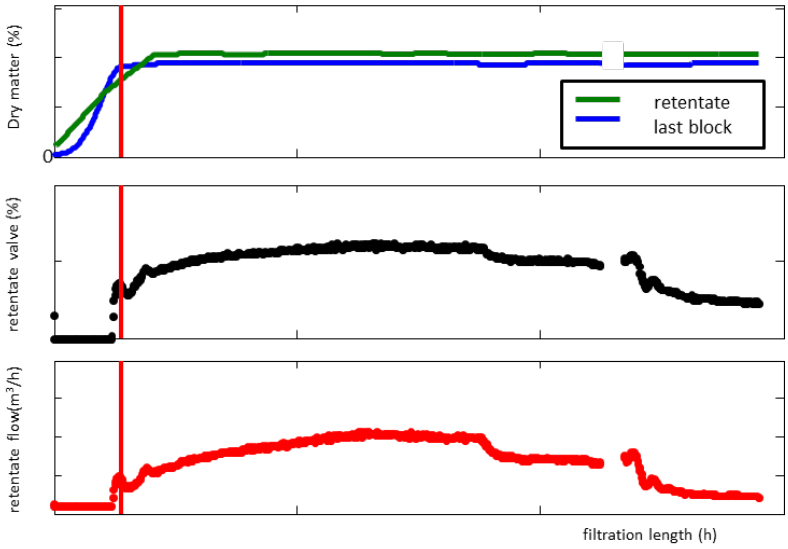


Figure 5-3 - Example of the separation of startup from a steady-state data. Data on the right side of the red horizontal line is considered a 'steady-state'.

The least troublesome procedure for initiating a multistage UF system is depicted in Figure 5-4. The two recirculation loops closest to the retentate outlet enter as the first as a product is fed to the system (Figure 5-4a). When the operator decides that the permeate flow is approximately correct in the last loop, the next loop counting from the end is added (Figure 5-4b). Additional loops are started until the plant operates at the correct capacity (Figure 5-4c-d). During a steady-state filtration (Figure 5-4c-d), the flow of the retentate and the retention time in the unit is regulated by the openness of the retentate valve. In the UF systems, initiating startup from the last loop helps to prevent excessively fast fouling. It is a well-known problem that the surge of high molecular weight components quickly blocks the membranes if the startup of the system is performed in the reverse order (that is by first starting block

A, then starting block B, and so on) [107]. The stages which were idle at the time of startup may not be necessary until many hours later. A decision on the number of blocks in use is primarily dictated by the capacity requirements of the process and controlled by the process operator. On average, the first blocks are the least frequently used but also other blocks can be switched off, e.g. due to technical issues. The last recirculation loop is normally always in use as the dry matter sensor used to control the unit is located there. In practice, this means that the structure of the data gets complex as different blocks are in and out off use for a different amount of time, or sometimes not used at all.

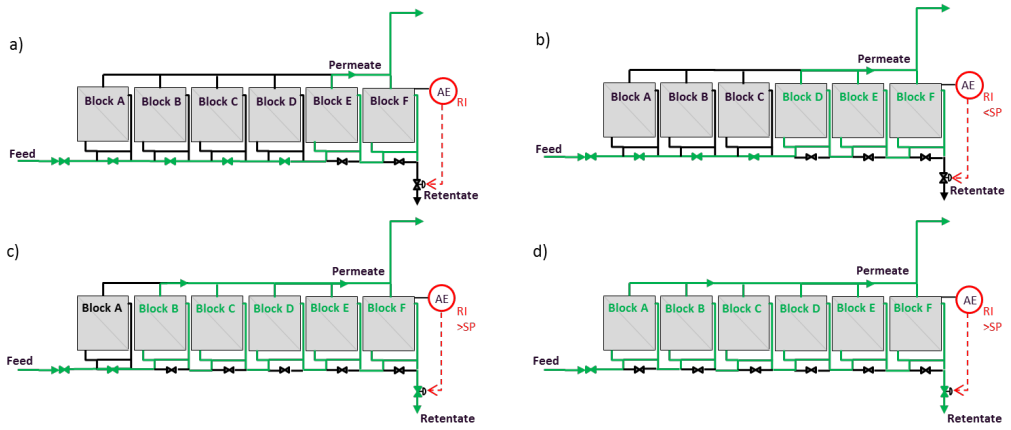


Figure 5-4 - From startup to the steady-state ultrafiltration in continuous processing: (a) initial startup with last two loops; (b) startup with three loops; (c) quasi steady-state filtration with five blocks; (d) quasi steady-state filtration with all blocks; AE-RI- dry matter sensor; SP – dry matter set-point.

Regulation of concentration degree

A membrane system designed as multi-stage recirculation plant with a high volumetric concentration ratio must be operated based on a very small flow of the retentate [107]. The actual control function is performed by a needle valve with actuator and positioner. The controller uses the sensor signals to set the desired position of the retentate valve. There are two possible manners to control the concentration ratio in the examined UF system. It is either based on: 1) the dry matter (refractive index, RI) measurement on the last recirculation loop; or 2) the volumetric concentration degree (VCD = feed flow in/retentate flow out). The first one is the

most common (presented in Figure 5-4). The main drawback of this control mode is that any precipitation of solids happening on the last block disturbs the refractometer readings. Because a powerful light source is needed to obtain a signal that can be detected, a local heating of the product could take place, and this can cause the product to precipitate and adhere to the prism of the RI sensor [107]. It is even worse if the concentrated product itself precipitates because then the system starts to work in a 'vicious circle'. A refractometer can measure only dissolved solids, so it does not see crystals and as a consequence concentrates even higher. For a product with this tendency, it can be a better practice to start up the unit using the dry matter controller but when the set-point is reached (that is during the steady-state) the control should be switched to VCD mode. If the composition of the feed is fairly constant, the former control function is satisfactory.

Real-time monitoring of enzyme activity

Relatively expensive process analyzers such as near infrared (NIR) instruments are increasingly considered for supervision of the quality and efficiency of industrial operations. NIR spectroscopy is suitable for timely measurements in dynamical systems, as the spectrum can be obtained quickly, without sample preparation, in a non-destructive way [100,109,110]. The potential of NIR technology to monitor the activity of the enzyme has been the subject of a feasibility study carried in the recovery factory at Novozymes A/S in Kalundborg. The work presented in PAPER I focuses on the UF retentate process stream (Figure 5-5). Internal studies showed that there is a strong relation between NIR spectra and enzyme activity in the UF retentate. Employing NIR spectroscopy, the indirect or inferential parameter enzyme activity could be obtained much faster than traditional off-line analysis in a central laboratory. It was possible to develop satisfying calibration models for four types of enzyme products. However, as it was desirable to develop a robust calibration, four QC parameters for enzyme activity have been standardized into one global QC parameter according to the formula:

$$Total\ Enzyme\ Protein\ (\%) = f_{Rescale}\left(\frac{Enzyme\ Activity\left(\frac{Unit}{g\ sample}\right)}{Specific\ Activity\left(\frac{Unit}{g\ pure\ protein}\right)} \times 100\%\right) \qquad (5\text{-}1)$$

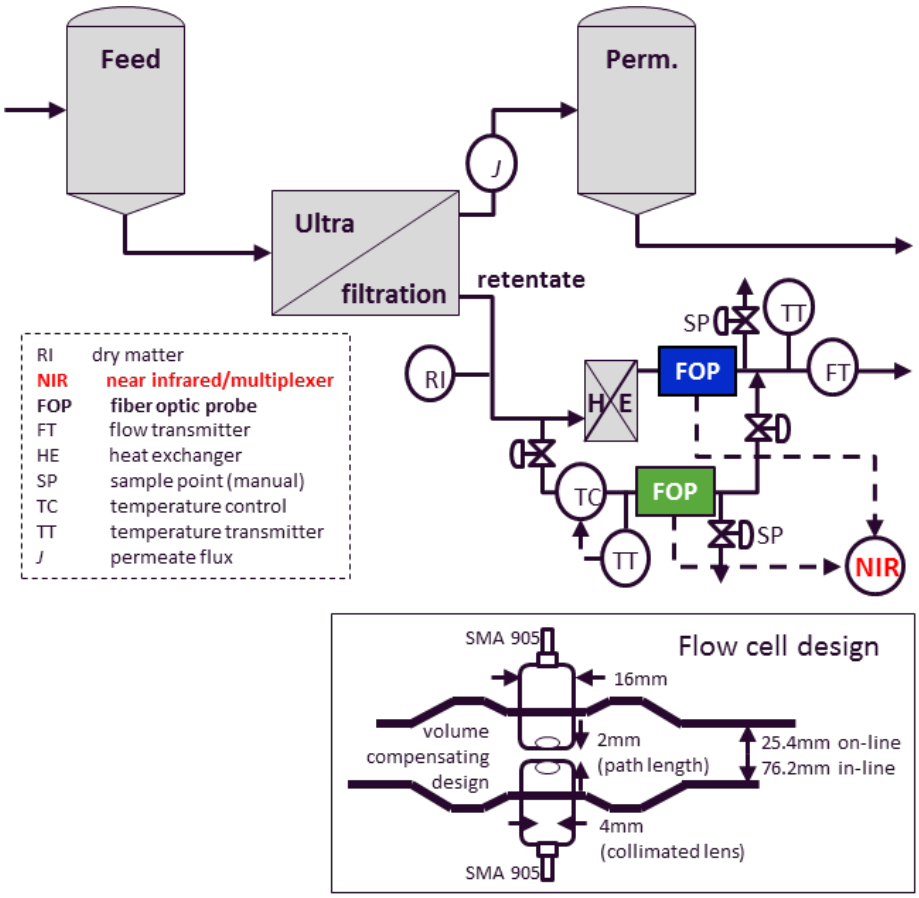


Figure 5-5 - Location of the NIR flow cells with respect to ultrafiltration unit. Both NIR flow cells – tuned to different dimensions of the adjacent pipelines, have the same design and optical path length.

When developing an on-line analyzer facility for determination of a multitude of process parameters, the challenge is not only to choose the right equipment, but also how to install it [3], and which quality control strategy to apply [103]. The NIR spectrometer uses optical fibers which offers the advantage of the multiplexing and eliminates the necessity for complicated sampling systems to bring specimens from the process to the analyzer. The probes can be directly mounted into the process line (called 'in-line' analysis, indicated in blue in Figure 5-5) or introduced in a fast loop with conditioning system ('on-line' analysis, indicated in green in Figure 5-5). The study presented in PAPER I involves (a) evaluation which of the two real-time NIR flow cells is the preferred arrangement, and (b) if the system can be used for statistical process monitoring and early warning/fault detection.

Both the in-line and on-line predictions deliver good, authentic results when compared to laboratory data. NIR predictions offered a sensitive and high-frequency feedback on the performance of the ultrafiltration operation. For monitoring purposes, it is sensible to keep track on the (spectral) variance not explained by the regression model (Q-residuals). Based on the PLS calibration step multivariate spectral information is turned into univariate predictions. Therefore, the cyclic pattern captured by the Q-residuals during a steady-state UF (Figure 5-6) is related to something else then enzyme activity. Interestingly, the fluctuations did not reveal themselves in any other (conventional) process measurement. Consequently, it does indicate the potential of real-time measurement as a tool in process identification and optimization next to concentration predictions.

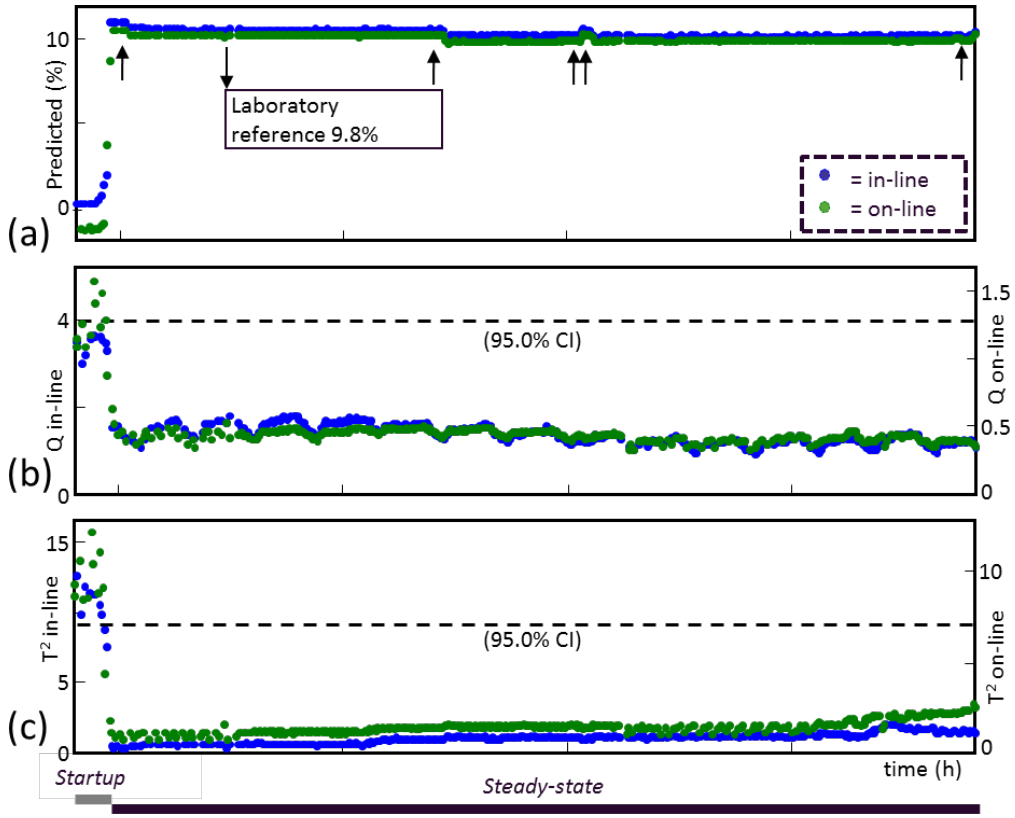


Figure 5-6 - Comparison between on-line and in-line NIR predictions of outcome downstream of the UF unit for a protease type product D; (a) predicted enzyme concentration, (b) spectral Q residuals; (c) Hotelling's T^2 score. Notes: arrows indicate operator set-point interventions on the concentration factor.

Finally, the study revealed that the less demanding in-line flow cell setup outperformed on-line arrangement. The former worked satisfactory robust towards different products (amylases and proteases) and associated processing parameters such temperature and processing speed. The disadvantages of the on-line setup include: more complex control of the sampling and conditioning system, and acceleration of the phase transition phenomena/sedimentation in the loop. It can furthermore be concluded that the method for unifying reference values from different analytical methods worked well.

Feed pressure control

Pressure is the driving force for membrane filtration systems. However, it cannot be used to control the capacity of UF plants. The capacity results from the nature of the feed and characteristics of the membrane. The operating pressure is factually irrelevant [107] as long as it remains constant at the predefined optimal value. In UF systems, feed pressure is controlled to guarantee that it does not exceed the allowed limits which could hurt the membranes. It is regulated by the power of the feed pump placed before the UF blocks (Figure 5-2e). The pressure decrease is the result of feed passing through the elements. Therefore, each recirculation loop is equipped with its centrifugal pump. The inlet feed flow is only indirectly controlled in the UF systems. It is the resultant of the retentate valve regulation and the pressure regulation in the unit.

In the examined system pressure is monitored at four points after the feed pump, one of which is also used in the closed-loop feedback control. Since pressure in the system is maintained very accurately, any deviation around the set-point is mostly noise. There were, however, particular periods across the examined production years when the set-point for the feed pressure had been changed. Those runs formed a clear cluster in the multilevel exploratory study presented in PAPER II. Nonetheless, it was suspected that the importance of the pressure set-point changes between the runs to the overall variance in the data had been unnecessarily blown up by the variance scaling. Moreover, an examination of the loadings plot of the within-model verified that there was no systematic variation to model in the pressure group. Consequently, it was decided to exclude three of the pressure variables to lower down the contribution of the pressure group to the model. In the following multiblock regression study (PAPER III), we have focused on one specific product

variant which faced a particularly steep flux decline. In this approach, we have identified that two out of forty runs exhibit a steep decline in pressure at all measuring points except the one used in the control loop. A new PLS model was calculated after iterative exclusion of (ten) abnormal runs and analyzed as a multiblock model. The study found that for investigated processing recipe the feed pressure to the unit showed a decrease over filtration run time, and it correlates positively to flux decline. Interestingly, except for the process tag directly used in the closed feedback control, the other pressure monitoring points show a drift over the filtration time. This observation would be impossible to make without limiting the investigation to one processing recipe and just by looking at the raw data before the abnormal runs were removed. 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.

Cross-flow over the membrane area

The UF system works to secure a sufficiently high cross-flow over the membranes in the recirculation loop. Feed flow per block is not measured directly but indirectly as the power consumption of centrifugal pumps. The cross-flow control loop assumes/uses the relation 'the higher the power consumption, the higher is the feed flow'. The single stage centrifugal pumps on the blocks run at a fixed speed. Throttling (butterfly) valves situated after these pumps (Figure 5-2d) get more open to compensate for the decrease in power consumption. What causes most of the power consumption is a liquid passing the outer edge of the impeller. The pumps and valves initially work at the predetermined settings ensuring the optimal, high cross-flow. Over a filtration run, a decrease in power consumption is typically observed (Figure 5-7). It indicates that less liquid is passing through the impellers so the feed flow decreases. The openness of the feed valve is regulated based on the power consumption of the booster pump. The feed valve opens wider to increase the flow and reaches the set-point for the power consumption. This regulation continues until the throttling valve is fully open. Afterward, only a decrease in power consumption is seen (Figure 5-7). This can be related to the increased viscosity of the pumped liquid.

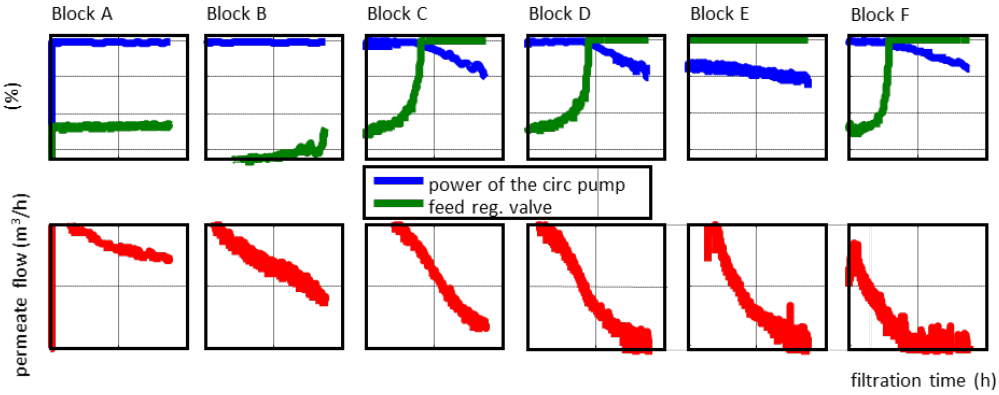


Figure 5-7 - Feed regulation on the blocks and corresponding permeate transverse flow values.

In fact, before the regulation valve fully opens, it is impossible to spot any changes in the power consumption based on the archived data. This is a consequence of the dead bands settings (see Paragraph 2.2). The power consumption momentarily recovers (faster than the pooling frequency for the compression in the DCS) as the feed regulation valve opens. This results in the significant differences between the effective compressions of the three signals in Figure 5-7. It is not feasible to examine the relation between the feed regulation on the blocks and the permeate flow/flux in too much detail. Though, it is indicative in Figure 5-7 that the cross-flow decreasing in the recirculation loops is a consequence of membrane fouling and not vice-versa. We identify this causal/mechanistic relation in the multiblock study presented in PAPER III. The highest variation in the flux is explained by the first LV, which had been predominantly linked to the cross-flow regulation. Therefore, an expert insight is necessary to clarify which of the observed relations can be used in optimization and which are expected from the mechanistic understanding of the system.

5.1.3 Data Alignment

Ideally, if retention times and/or lags are identified, this information can be used to properly align data from subsequent operations in a continuous process. Namely, one row or time-point in the data matrix **X** should ideally contain information corresponding to the same entity of the processed material, and it should have the corresponding measure(s) of the final quality assigned to it in the response matrix **Y**. In other words, all the values for the process parameters assigned to one row should

correspond to a certain end-quality when applying the linear algebra methods from chemometrics.

With respect to process tags in a continuous UF operation, it is natural to think of 'lagging' in relation to the physical distance between the sensors. E.g., it seems unnatural to compare at the same timestamp the parameters measured before the UF unit (in the feed) with the parameters measured after the UF unit (in the retentate and permeate) due to retention time in the UF unit. If two time series are shifted so that they are offset in time, the potential correlation between the two parameters can be missed. Therefore, in a seamless approach, each row should contain parameters corresponding to the same part of the (original) feed. However, in the investigated unit operation, the correct lags are almost always unknown. In general, it would be extremely hard to establish even the retention time in the UF unit because it would vary continuously owing to a different number of the blocks in use, the degree of recirculation on the blocks, process temperatures, properties of the feed and the concentration degree. Some trial-investigations on this have been perfoemd during this Ph.D. with tools such as serial correlation analysis and PCA on the lagged dataset followed by observation of loadings. However, these studies did not provide any hints with respect to suitable lags. One reason for this could be the varying logging frequency from different process sensors. Some process parameters are close to steady during the entire run (e.g. pH, dissolved solids, temperatures). Hence, it is not expected that shifting the signals to match within a minute precision would make any difference. Tags that are logged with a high frequency such as flow tags, and related regulation valves, are conjugated via the control loops. In this case, the highest correlation always appears at the same timestamp (lag zero) because the response to the regulation impulse is observed in all of them at the same time. Since it is so complex to track the path of a product/effluent stream in this UF operation, it was decided not to align the signals from different sensors in the UF studies. Instead, the average values over a fixed and equidistant time interval were used.

UF runs cannot be stacked on top of each other to form a three-way matrix in a straightforward manner as they are not truly equal. A different length of the runs is dictated by varying volume and parameters of the feed and different set-point parameters of the retentate and the number of blocks that are in use. The volume and parameters of the feed differ considerably depending on the conditions and

performance of upstream operations (fermentation, separation, pre-filtration). For that reason, it is impossible to point to some indicator variables which characterize a common start and end point for all runs. If the runs were equal, then it would be expected that the calculated variable 'mass throughput' reaches the same or similar value at the end of each run, but it never does (Figure 5-8).

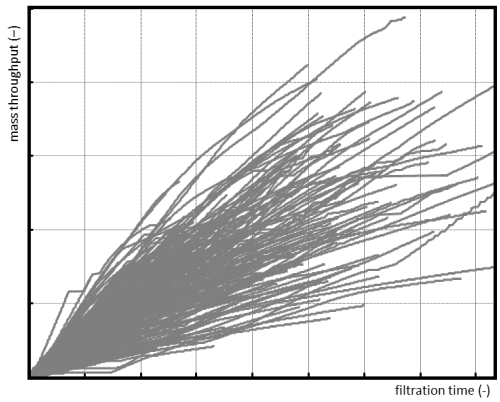


Figure 5-8 – Overview of mass throughput for 278 runs examined in a UF study.

The process runs examined in this study are characterized by different mass balances. In other words, those runs cannot be considered as replicates of each other since they are collected under quite different processing conditions. Moreover, a steady-state ultrafiltration process cannot be perceived as a multistage process suitable for synchronization as there are no natural 'phases' or other systematic information that could be used to synchronize towards. This should clarify that the studied data is not suitable for synchronization by any of the approach discussed previously (Section 3.4.2 'Synchronization'). Those methods might correct the start and end phases of a semi-batch/semi-continuous process, but the startup data has been excluded from the current investigation.

In the end, it was decided that it is only possible to arrange the data into the three-way structure by limiting the data to a fixed length (Paragraph 5.1.5). Namely, the runs have been analyzed up to the median length for the dataset consisting of two hundred seventy-eight runs. By excluding shorter runs this leaves one hundred forty-one cases. This way, it is expected to get a better idea about the stability of the correlation structure over the course of ultrafiltration using the BWU methods (Figure 4-4 F). Alternatively, one could consider limiting the data to the median mass

throughput value of the process. This was attempted, but this operation did not lead to any revelations and made models more complex to interpret.

5.1.4 Examination of the flux decline

Ultrafiltration can run only for a limited period before the membranes have to be cleaned [111,112]. The UF capacity is limited by a membrane fouling or blocking phenomena where the production capacity - monitored as flow through the membrane, or flux - decreases over time. In general, the flux decline is caused by a decrease in driving force and/or increased resistance [113]. This fouling mechanism has been studied for the past 40 years [113-117]. It is obviously very complex, but main attention was focused on the nature of the membranes, the parameters of the processed feed and the processing conditions. Normally, the flux is much higher at the startup of the process, which corresponds to the pure water permeability of the membrane. After a few minutes, the flux sharply decreases due to the concentration polarization (represented with black markers in Figure 2-4). The permeate flux will after that continue to decline gradually due to membrane fouling (represented with red markers in Figure 2-4) [118]. Concentration polarization involves an increased concentration of retained components close to the membrane surface and it depends on the hydrodynamic conditions in the membrane plant [23,113]. This part of the flux decline is reversible and can be reduced for instance through flushing of the system with water, increasing the cross-flow velocity, decreasing the concentration of the feed or decreasing the transmembrane pressure [105]. As a result of the concentration polarization at the membrane surface, increased ion concentrations and other feed components can surpass solubility thresholds and precipitate on the membrane surface and/or in the pores of the membrane [117,119]. These deposits can block pores causing a loss of performance. As it was decided to exclude the startup data from the capacity studies, the examined flux decline is predominantly related to flux decline which ideally should be recovered by cleaning. The CIP-irreversible blocking of membranes might also occur [111,112], but it is not in the scope of this study, due to lack of available data on 'pure water flux' after cleaning.

Understanding run-to-run variations in performance of ultrafiltration in terms of flux has been the scope of PAPERS II, III and POSTER II. One of the preliminary challenges associated with those studies was to establish a suitable estimate of the flux for the post-run evaluation of UF performance. In daily practice, the UF capacity

is monitored based on the permeate flow out of the UF loops and the retentate flow. The operator stops the unit and proceeds to cleaning when these parameters drop to unacceptably low values or if the order is finished. It is, however, problematic to use the seven parameters for the post-run capacity evaluation, especially since not all UF loops are in use all the time. Moreover, runs significantly vary in duration and volume of the processed feed. Naturally, the higher is the feed volume, the longer is the processing time (Figure 5-9a). However, it can also be observed that as the feed volume increases the variation in the processing time between the runs also increases. Consequently, the UF performance can be expressed as a feed volume processed per unit time (Figure 5-9b).

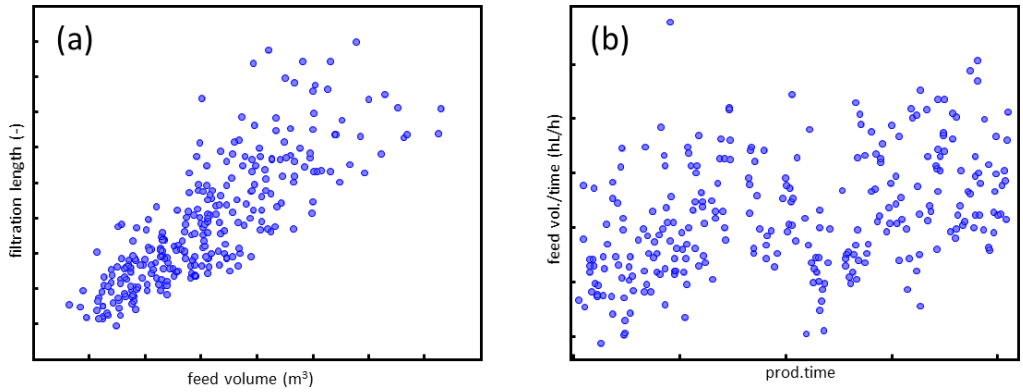


Figure 5-9 – (a) Relation between the processed volume of the feed and the run length; (b) overview of the UF capacity expressed as feed volume per unit time.

This very crude way to address the problem translates to capacity in a straightforward manner but has several drawbacks. For instance, it does not compensate for the working membrane area during the run (that is a number of UF blocks in use). On the other hand, process engineers normally evaluate the post-run performance based on the appearance of the permeate flow trajectories for the entire run. For instance, Figure 5-10 presents permeate flow profiles for six subsequent UF loops experienced during three different runs.

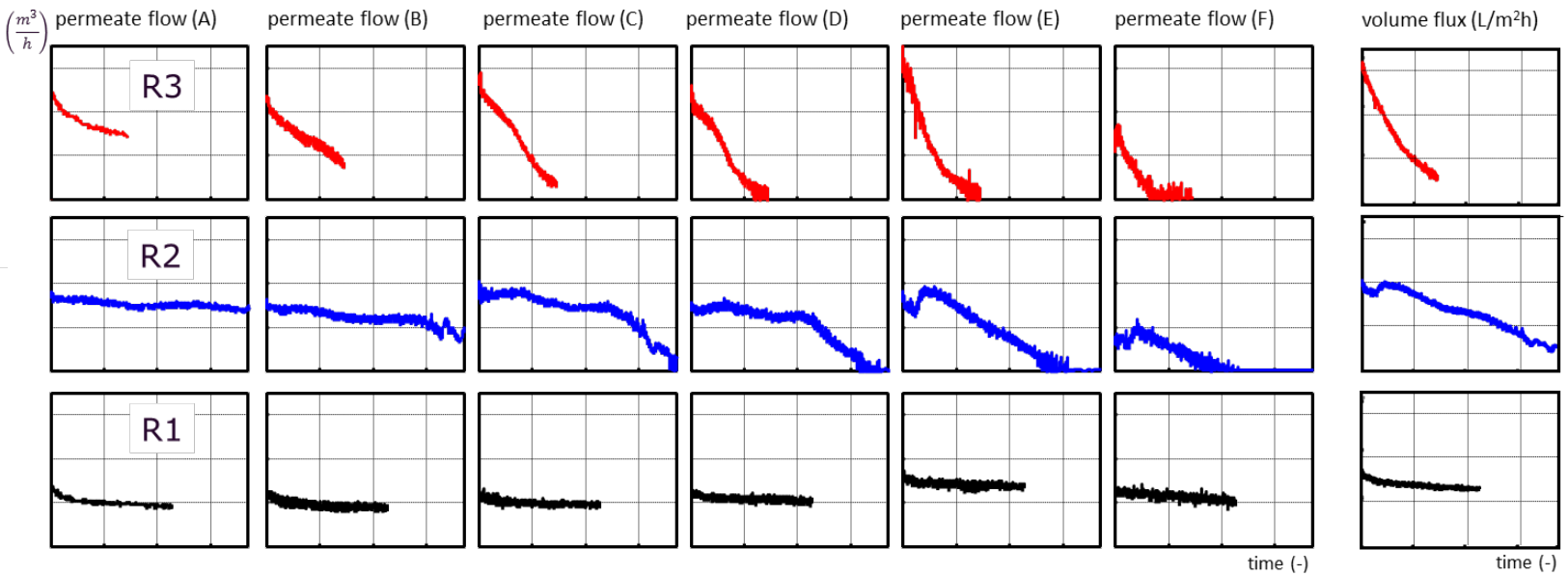


Figure 5-10 – Examples of three permeate flow profiles (R1-3) per block (A-F) corresponding to quasi-steady-state filtration and the volume flux calculated from them according to Eq.(5-2).

Once again it can be noted that the filtration duration varies considerably between the runs. A common observation valid for all examined cases is that the decrease in the permeate flow appears first on the last recirculation loop F. This UF stage is always in use; it enters first in production, so it also works for the longest time. Recirculation loops work in sequence and there is thus a successive increase in total dissolved solids moving from stage A to F. Consequently, the last stage is the most exposed to fouling. Flux decline is always the most severe there, and the problem transfers to the other blocks later in the process in reverse order from F to A. In the above example, run R1 represents a desirable profile. The permeate flow on the last block is decreasing steadily but very slowly. As some decrease is practically always observed, it is more correct to refer to the filtration phase, which follows after startup, as a 'quasi-steady-state'. Both runs R2 and R3 represent bad profiles. According to the experts, run R3 is the worst case scenario where the permeate flow decrease very steeply right after startup, which is observed on all blocks. On block F, permeate flow reaches zero already mid-way through run R3 and the preceding blocks soon follow. As a result, this run is very short. Run R2 also experiences a steep decline in permeate flow on the two last blocks, but it is less severe than in the previous case. All in all, this run is 2.5 times longer than run R3. The expert evaluation is reflected in Table 5-1 by rating the best run R1 as 1st and the worst run R3 as 3rd. The subsequent challenge is to reproduce the experts' assessment based on potential flux estimates/parameters calculated from available data, as listed in Table 5-1.

Table 5-1 - Flux estimates corresponding to Figure 5-10. The ratings in each column are based on the assumption that the higher the value, the better was the UF performance.

Run	Expert assessment of profiles	permeate flow, mean value ($m^3.h^{-1}$)							feed vol./time
		block A, v_A	block B, v_B	block C, v_C	block D, v_D	block E, v_E	block F, v_F	total v_{tot}	
R3	3rd	1st	1st	1st	1st	2nd	2nd	1st	1st
R2	2nd	2nd	2nd	2nd	2nd	3rd	3rd	3rd	2nd
R1	1st	3rd	3rd	3rd	2nd	1st	1st	2nd	3rd

None of the approaches listed in Table 5-1 can reproduce expert evaluation. This is because when using means one disregards the starting value, the slope and the filtration duration. Run R3 does not appear as the worst according to any of the investigated parameters. It is the one most frequently scoring as the best. Only mean permeate flows on the last two blocks position run R1 as best. Surprisingly, comparison of the post-run capacity expressed as a feed volume processed per unit time results in the opposite ratings than in case of the experts' assessment of the profiles. Additional problems, not covered in the above example, appear when not all blocks are in use. This, however, can be overcome by calculation of 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:

$$J(t) = \frac{v_{tot}(t) \cdot 1000}{Wb(t) \cdot A} \qquad (5\text{-}2)$$

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 number of loops working at timestamp t, based on assumption that a loop is working if the power of the corresponding centrifugal pump is larger than 1%, and A is membrane area corresponding to one loop (m^2).

The fluxes calculated according to Equation (5-2), corresponding to the previously discussed three runs, are presented as the last column in Figure 5-10.

The flux profiles elegantly summarize the information that has been previously evaluated per block. The most desired scenario is to have a steady flux profile as in the case of the run R1. Run R2 shows a flux decline but the performance is quite good until approximately midway through the run. Run R3 has an exceptionally high flux at the start but shows a drastic flux decline over the ultrafiltration. The flux at the end of the run R3 is comparable to the end flux of the run R2 which was 2.5 times longer. Capacity wise (feed volume/time), run R3 outperforms run R2 but in the case of the latter twice as much feed had to be concentrated. If the feed volume in R3 was equal to the feed volume in R2 it would be impossible to finalize the process order without an additional CIP which would be an unexpected capacity cost. Consequently, it is not an easy task to decide on a universal post-run capacity estimate in a continuous ultrafiltration process that satisfies all nuances of economic

production. Therefore, it was decided that flux will be parameterized by three numbers (Paragraph 5.1.5):

1) Start value – mean value for the first 30 min of the steady-state filtration.
2) Mean value – overall mean flux value for the steady-state filtration run.
3) Slope – the difference between the start value (first 30 min) and the end value (last 30 min) divided by filtration length.

The calculated values corresponding to the discussed examples are rated with respect to each other in Table 5-2.

Table 5-2 - Flux estimates corresponding to Figure 5-10. The ratings in each column are based on the assumption that the higher the mean/start value the better, whereas the higher the slope value the worse was the performance of the UF.

Feature / Run	Start value ($L/m^2/h$)	Mean value ($L/m^2/h$)	Slope ($L/m^2/h$)
R3	1st	1st	3rd
R2	2nd	3rd	2nd
R1	3rd	2nd	1st

The flux estimates comply well with the experts' evaluation of the profiles, and they are easy to understand and translate back. The slope component takes filtration time into account which is also good because it is obviously better if the comparable decrease in flux is extended in time. The high start and mean value for R3 corresponds to the high capacity which is paid for by a fast flux decline (slope). It appears wise to investigate the relation between other process information and these three values at the same time, and chemometric modeling can handle the reference in this format (via PLS2). An overview of the variation in these three features in the investigated historical production runs (I = 278) is shown in Figure 5-11.

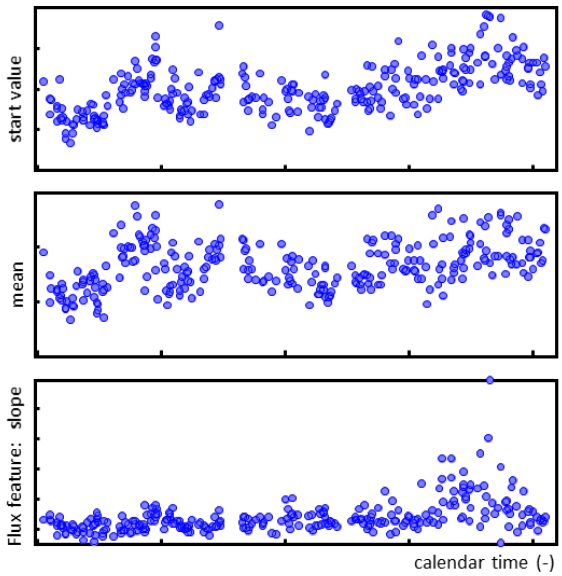


Figure 5-11 - Overview of the values for the flux features extracted for all available runs of the same commercial enzyme intermediate product as a function of time.

To summarize, during the UF capacity studies flux decline has been addressed from different angles to make the studies comprehensive. It is recommended to select the flux estimate that works best for the purpose of the study. Depending on the modeling approach reference flux has been examined as time series 1a) 'Y-flux' (volume flux calculated for all blocks) (PAPER II and III), 1b) 'F-flux' (permeate flow on F-block only), 2) translated to bins (POSTER II; Paragraph 3.5), 3) Y- or F-flux features or as 4) feed volume processed per hour. Options '3)' and '4)' have been studied in more detailed in a separate, internal report for Novozymes.

5.1.5 Modeling

The flux decline problem has been addressed throughout this project in several ways for one specific type of commercial enzyme product. At first, a flux quality map had been formed by the PCA decomposition of the binned 'flux on the F-block' (POSTER II). The position of the runs on the first PCs score clearly showed that those UF runs which require less up-concentration have a better flux behavior. Hence, they might not be so interesting to examine. It was not straightforward what causes the diversification across the second PC between the good and bad fluxes for the runs with higher degree of concentration in the retentate. Another observation was that the runs processed according to the newest recipe were situated either in the low flux

corner of the model or close to the center of the plot. This could indicate that they are not well explained and behave differently. In the following studies, the use of the binning for data was withdrawn and replaced with calculated flux and flux features. In the meantime, CCSWA (Paragraph 4.4 and Appendix 1) was used to examine the association between different process parameters and to support the data validity check and selection of the relevant tags. Originally, the historian call returns seventy-eight process tags related to the examined UF unit. Quickly, cumulative, CIP-related and operational tags were excluded which left sixty-six process parameters. After data validity checking and use of preliminary process knowledge forty-eight parameters were judged relevant for the PLS-based studies. As features extraction is the simplest and the most flexible of the considered data arrangements, it was used first. This approach enables for easy incorporation of upstream process parameters and other calculated variables (as summarized in Table 5-3). Seventy-three parameters expressed by one to three features (not all identified in Table 5-3 due to proprietary reasons) were investigated at this stage, adding up to one hundred forty-seven variables.

PLS2 models were calculated between the flux features and features of process parameters using all available runs, and six components were selected based on the scree procedure [120] (Figure 5-12a). This model is referred to as FE-PLS2. The model had been optimized by the VIP variable selection technique. It was possible to limit the number of variables to only fifteen without loss of performance, based on a stratified CV procedure (4 LV's, Figure 5-12b). Parameters which remained in the model were related to the temperature of the feed (predominantly), feed and retentate RI, and cross-flow regulation on the second and the last UF-loop.

Table 5-3 - Process tags and other calculated parameters investigated in relation to flux and expressed as features.

Unit operation (Figure 5-1)	description	How is it expressed?						
		mean	st.dev.	range	ratio	length	cumulated value	profile
Fermentation	fermentation activity	X						
Pretreatment	flocculation agent, flow	X						
	dilution	X						
	pH	X						
Separation	conductivity after centrifuge	X	X					X
Pre-UF	RI retentate	X						
UF	membrane age				X			
	filtration length					X		
	'idle' time					X		
	startup length					X		
	volume reduction factor	X	X				X	
	mass throughput						X	
	working block score				X			
	permeate return				X			
UF feed	pH	X	X					
	temperature	X	X					
	reg. pump feed tank	X	X					
	pressure 2	X	X					
	pressure 3	X	X					
	pressure 4	X	X					
	pressure 5	X	X					
UF per block only when in use (A,B,C,D,E,F)	start time				X			
	time in use				X	X		
	dilution mode				X			
	feed regulation valve	X	X	X				
	power of circulation pump	X	X	X				
	Temperature	X	X		X			
	cooling (ice water reg. valve)	X	X					
UF retentate	RI block F	X	X	X				
	RI retentate	X	X	X				
	temperature	X	X					
	cooling (ice water reg. valve)	X	X					
	Conductivity	X	X					X
UF permeate	pH	X	X					
	Temperature	X	X					
	Conductivity	X	X					X

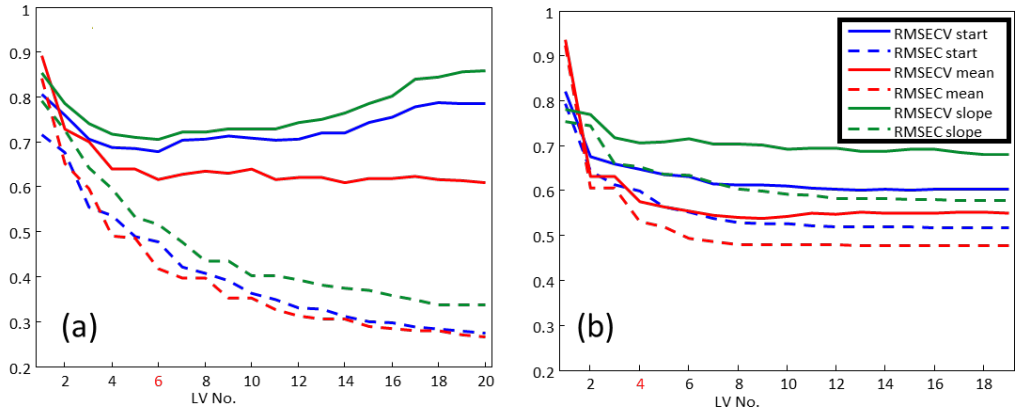


Figure 5-12 - Calibration (RMSEC) and cross-validation (RMSECV) errors for the FE-PLS2 models constructed using (a) all data (J = 147) or (b) VIP selected tags (J = 15).

Next, the dataset which consisted of runs limited to the median length (size 147×57×N_m) was studied by the BWU techniques. Figure 5-13a and b present mean trajectories of investigated variables plotted against the filtration time. The effect of auto-scaling on this data in a BWU vs VWU arrangement (Figure 4-4 F vs D) is compared in Figure 5-13c and d. Auto-scaling in VWU (so-called 'variable scaling + centering') does not remove the mean trajectories from the data which is the most prominent for process parameters such as conductivity, flow tags and cumulated values (Figure 5-13c). As a result of auto-scaling in BWU ('trajectory scaling plus centering'), means of all variables at every timestamp are virtually zero (Figure 5-13d). After BWU, a 2-D matrix of size 147×57N_m is obtained.

Next, a 'Total Covariance' map (see Section 4.6) has been consulted to visualize the dynamic relationships in the data [86]. Explicitly, the motivation for using the map was too see if the correlation between the investigated process parameters is time-varying. Again the effects of auto-scaling in the VWU and BWU arrangements have been compared. The two pre-processing strategies resulted in almost identical maps (except for cumulated values). Figure 5-14 shows the 'Total Covariance' map for the data scaled in the BWU arrangement. It is a visualization of the auto-covariances and lagged cross-covariances among process variables over time. The sampling time is indicated in the margins of the map. All examined variables are present within each block of each sampled time. Intensive colors indicate a strong correlation between

variables whereas white and light colors indicate a lack of correlation between variables and phases.

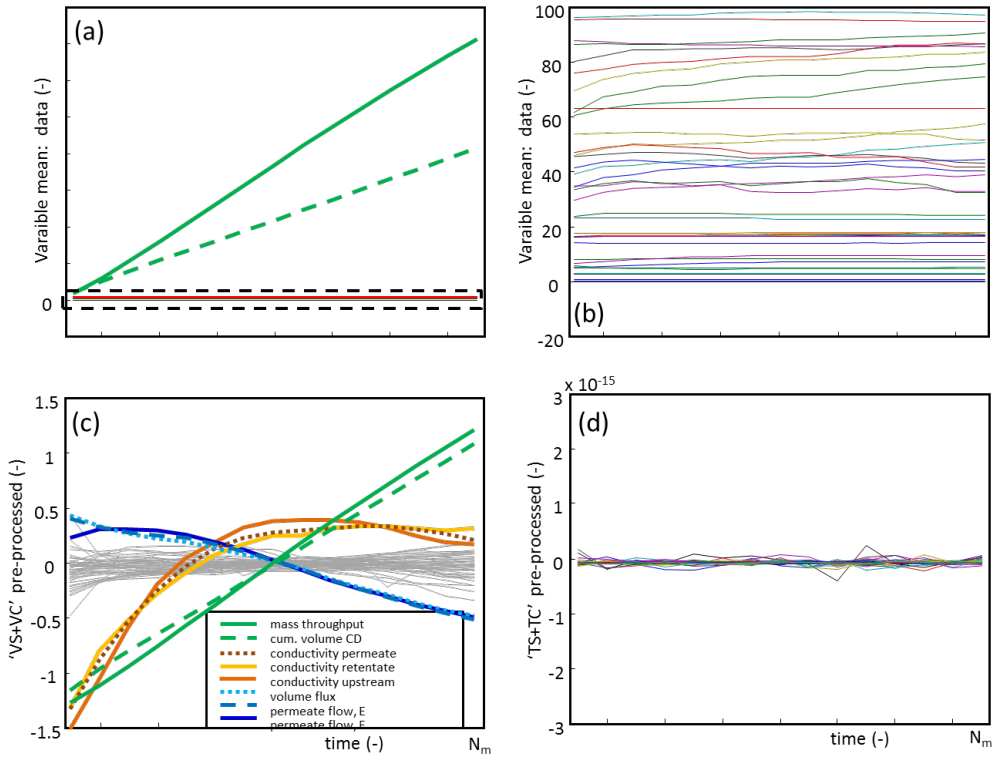


Figure 5-13 - Mean process variable trajectories used in BWU-PCA approach before and after scaling; (a) before pre-processing; (b) zoomed region indicated by broken line in (a); (c) after variable scaling plus centering; (d) after trajectory scaling plus centering.

Overall, judging from this map, the covariance structure among investigated process variables appear to be quite consistent with time. The intensive red squares between variables 6-10 repeating for all time lags represent the strong correlation between the pressure related tags. It appears that correlation between variables on the first three UF-loops (blocks A-C), which are situated on the map just after the pressure tags, is stronger in the beginning and diminish over filtration time. If a process runs through different phases, it should be possible to observe it in the correlation map as the formation of rectangles described by distinct color pattern [46]. However, it is not seen in examined UF data. This, together with the lack of notable difference between the auto-scaling in the batch-wise vs. variable-wise arrangement, suggests that

correlation structure in the data is quite constant. Therefore, the VWU-approach (Figure 4-4 D or E) will most probably work sufficiently well for the modeling of UF process data.

Similarly, there was little difference in the shape of the loadings between the BWU-PCA which used data after 'trajectory centering + scaling' and the BWU-PCA model based on data after 'variable centering + scaling' up to the third PCs, except for the cumulative values (not shown). The summary statistics of the two models were very similar as well as the appearance of the BWU-PCA scores. Consequently, the model interpretation would be the same irrespective of the scaling performed.

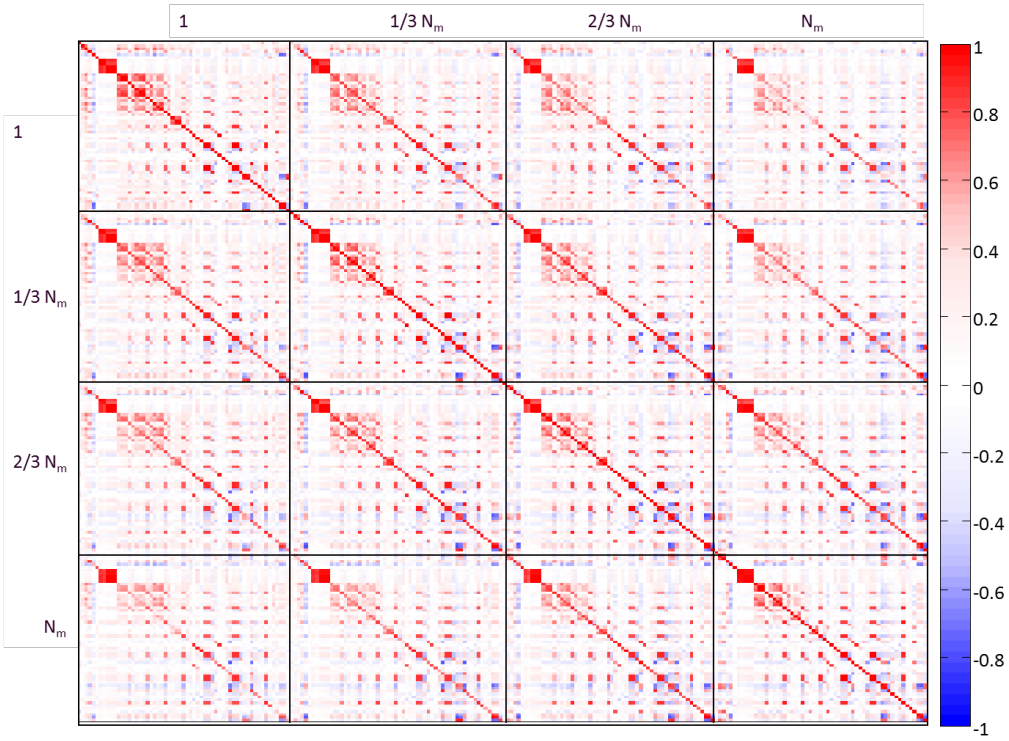


Figure 5-14 - Correlation map for the batch-unfolded and auto-scaled data (147×57N_m) for ultrafiltration hours: 1, 1/3 N_m, 2/3 N_m, N_m. Fifty-seven process tags are arranged in an order which corresponds to the UF process layout/effluent flow.

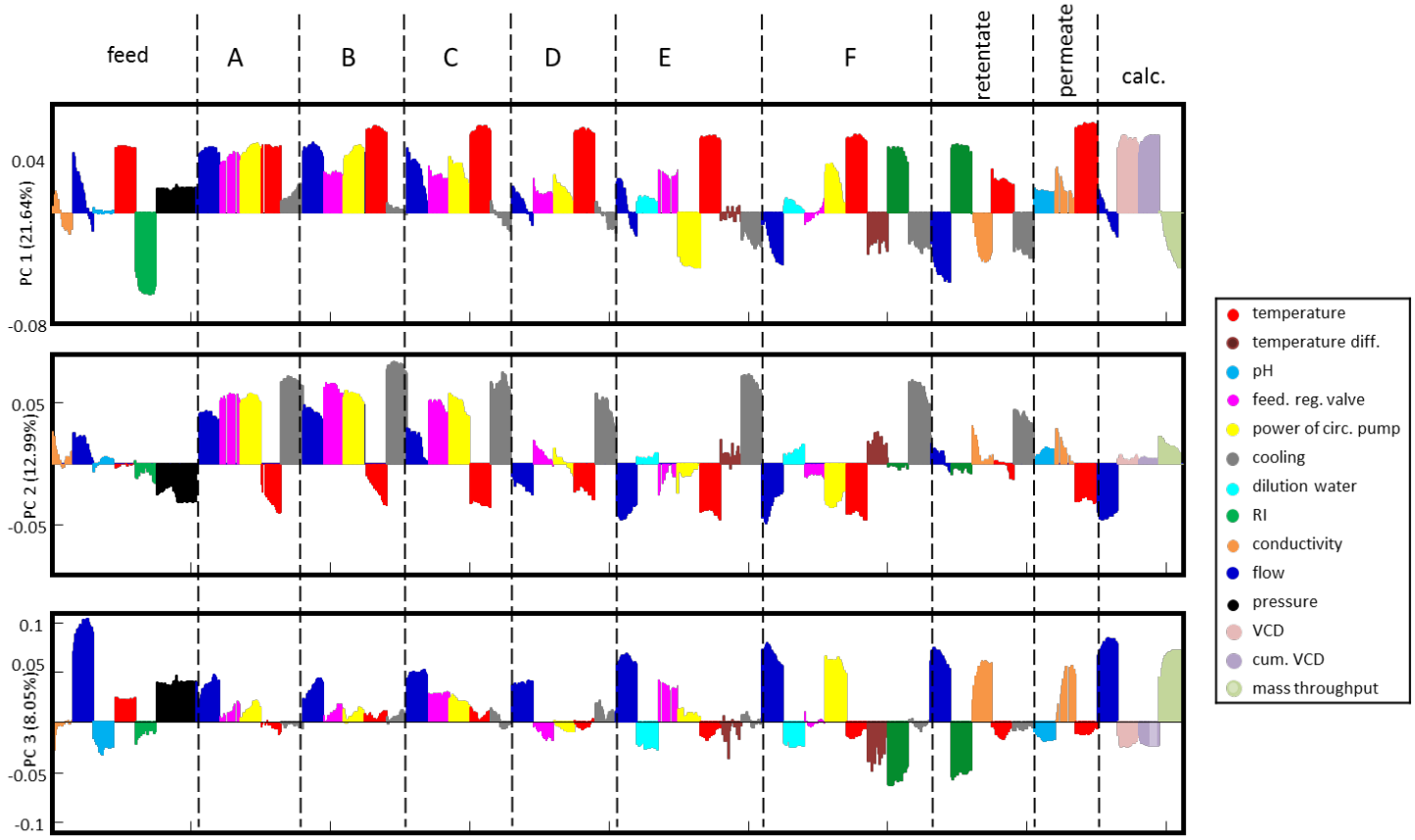


Figure 5-15 – Loadings of the final BWU-PCA model (3 LVs, X: 146 × 54 N_m). Every process tag is represented by N_m loading values plotted from time one to N_m. Loadings bars are colored according to the sensor type.

The motivation behind the use of BWU-PCA in this UF study was to check if modeling of the flux decline benefits from the methods capable of handling the changing process dynamics. From the shape of the loadings (Figure 5-15), it can be concluded that the correlation structure does change slightly. For instance, when a UF-loop is 'OFF' (which is usually the case for the first UF loops at the first timestamps) the temperature is equal to ambient temperature of the surrounding air. Thus, no correlation between the cooling regulation and the temperature should be observed for this stage. If the loop is working then, in general, the more intense the cooling, the lower is the temperature (but also the reverse situation could happen). Finally, if the cooling reaches its maximum capacity the temperature may still increase, especially if the membranes get more and more obstructed. In this case weak or no correlation should be observed. This situation would be primarily expected on the last UF loops at the later timestamps. This 'control loop' and 'ON/OFF' blocks dependencies can be the reason behind the varying shape of the loadings on the second PC. The profile of the flow tag loadings on the third PC also make sense as we observe an increase in loading value over run-time for blocks A and B (blocks enter later) and decrease for F-block and retentate flow (fouling, lower throughput). The flux and feed flow shape are a mixture of the other flow tags to which they are related via the closed-loop control or, in the case of flux, calculation. All in all, interpretation of the shape of the loadings obtained with BWU-PCA is not easy or straightforward. It would not be possible without a solid process experience and it does not lead to any new findings.

It is doubted if the BWU data arrangement is needed to examine the UF flux problem. The process does not run through fixed phases. Hence, it cannot be predicted beforehand when a specific block or control loop is active, and it is thus not possible to define an NOC set accordingly. A more repetitive nature of the process is necessary if a BWU-PCA-based monitoring/optimization strategy is desired. The shape of the BWU-PCA loadings obtained in this study just reflects the averaged UF performance. Moreover, when it came to the investigation of outliers (not shown), it was never the shape of the loading that indicated the outlying behavior but rather the entire block of variables being off. The BWU-PCA and PLS2 (after exclusion of flow tags) models has been examined in more detail in a separate Novozymes internal publication. The BWU-PLS2 model after variable selection used similar process parameters as a FE-PLS2 model.

For comparison, a multilevel PLS model has also been developed based on the limited dataset (size: 147 N_m × 46). This regression approach uses the flux series as an independent variable, and it models the relation to the flux on the 'between-run' level and 'within-run' level separately. The performance of the optimized multilevel PLS is summarized in Figure 5-16. Markers correspond to different processing recipe variants of the same commercial product. The between-runs model predicts the mean flux value (I = 147) whereas within-runs model predicts the variation in flux over the course of ultrafiltration (for 147 N_m timestamps). The latter sub-model is comparable to the start and slope features used in FE-PLS2. The optimized multilevel PLS model used eighteen variables. The remaining variables had different importance in the sub-models as judged from the corresponding VIP-scores (not shown). For instance, feed temperature was primarily used by the between-level sub-model, whereas variables related to the cross-flow regulation on the UF-loops were more significant for the prediction of 'within-run' flux variation. The relation to conductivity was also relevant on the within-run level. However, this could be driven by a similar development of the flux and conductivity profiles over filtration, just reversed. Explicitly, conductivity normally increases over the course of the UF process, and flux decreases (see Figure 5-13d).

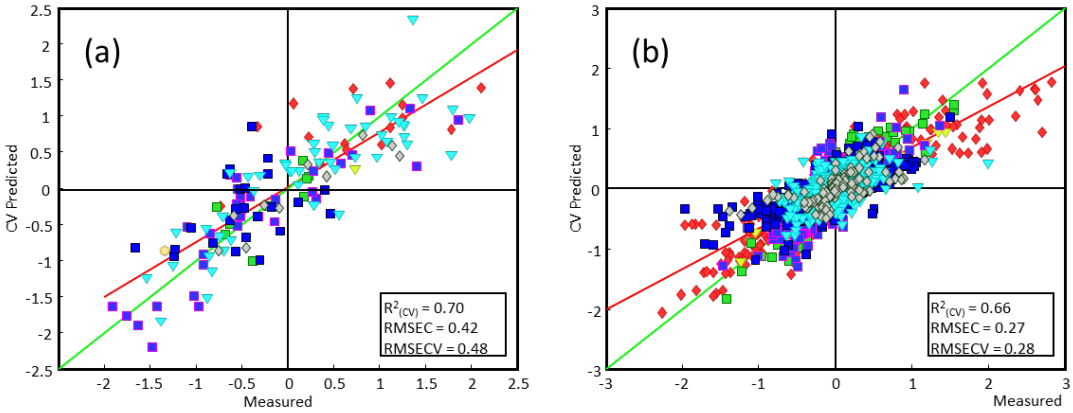


Figure 5-16 – Performance of the optimized multilevel PLS sub-models (a) between-level (6 LVs); (b) within-level (3 LVs). Flux is expressed in units of its standard variation.

Processing recipes marked with red rhombi show the highest spam of flux variation within the runs (Figure 5-16b) and the highest overall mean values (Figure 5-16a). In the case of runs marked as turquoise triangles and gray rhombi, flux does not vary

much over the course of filtration (values close to zero in Figure 5-16b). The lower fouling potential is expected for these recipes as they demand a lower degree of concentration.

A detailed description of the outcome of the exploratory (MSCA-P, BWU-PCA) and regression studies (FE-PLS2, BWU-PLS2, multilevel PLS) in the context of the flux decline cannot be included in this thesis as it would be hard to follow by an external reader. All approaches pointed at the relation between the flux and the feed temperature where in particular high RI products experienced a steep decline in flux. This observation was already clear from the conceptually simplest modeling strategy applied to the broadest dataset (FE-PLS2 for 278 runs). The multilevel exploratory study (PAPER II) focused only on the high RI variants of the product (141 runs) and has the big advantage of maintaining the time series structure in the examination. The underlying factors behind recipe-, between- and within-levels of the MSCA-P model have been identified, which greatly eased the exploration of the fifty-seven parameters used to monitor the process. The dominating phenomenon in the within-run variation has been related to the throughput/flux profiles. Runs processed according to the most recent recipe showed distinct profiles which corresponded to significantly higher throughput at the start followed by a very steep decline. The steep flux decline as well as higher overall mean flux value appear to be related to processing temperatures and, in particular, increased feed temperature.

The flux decline problem has been further addressed by the MB-PLS approach (PAPER III) centered only on the last recipe variant. In the study thirty runs has been classified as NOC and ten as AOC based on their behavior on LV1 vs. LV2 score plot. Similarly to our previous study, it appeared that higher processing temperature can have both positive and negative consequences for the UF flux. Specifically, it was found that the extent of membrane fouling can be reduced if the temperature on the last three recirculation loops is decreased. The reason for rapid blocking of the membranes was most probably the enhanced phase transition of the enzyme or precipitation of salts which for this particular product happens faster at high temperatures. Membranes 'do not cheat' and are designed to reject the suspended solids hence precipitation in the unit should be prevented at all costs [107]. Additionally, in the last study a potential field for improvement has been reported, related to the pressure monitoring point used in the closed-loop feedback control of ultrafiltration pressure.

5.2 Production of enzyme granulate

The solid formulation of enzyme into the granulate product rather than into fine powder is motivated by the reduced risk of the airborne particles. Strong and resilient granules ensure a 'dust-free' product which is safe to handle in the clients' factories [121]. Other advantages of this solid form are improved flowability and homogeneity, reduced risk of segregation and better stability during storage. Formulation of dry enzyme products at Novozymes A/S in Kalundborg is performed by 1) continuous granulation and drying followed by 2) batch coating and cooling process (Figure 5-17). The wet granulation process is carried out in high-shear mixers using the mechanical energy of chopping blades and plows which distribute the liquid binder.

The sub-processes of wet granulation are: 1) wetting and nucleation, where a granulation liquid (enzyme concentrate and water plus dextrin) is sprayed over the blended powders (fillers such as cellulose, dextrin and various salts); 2) growth and consolidation which involves the coating of fresh powder onto the surface of the granules and propagation in size through sticking two or more granules; 3) attrition and breakage during which the surface of the granules wears down progressively [122].

Next, the wet granules are dried in a fluidized bed. After drying, the raw granules are sieved, and the oversized and undersized fraction is recycled in the process. The raw granules of appropriate size are coated in order to keep the dust level to a minimum and to prolong the stability of enzymes. One or more layers are applied, and the coating is usually an inert material like organic polymer and/or salt. The coating can also be used for coloring purposes. For instance, titanium dioxide can make the product whiter [121]. The physical strength of a granule and the coating applied to granules is an important quality parameter to monitor when granulating enzymes. Enzymes are allergens, and hence, it is vital to secure that none are released to the surrounding environment due to poor coating quality.

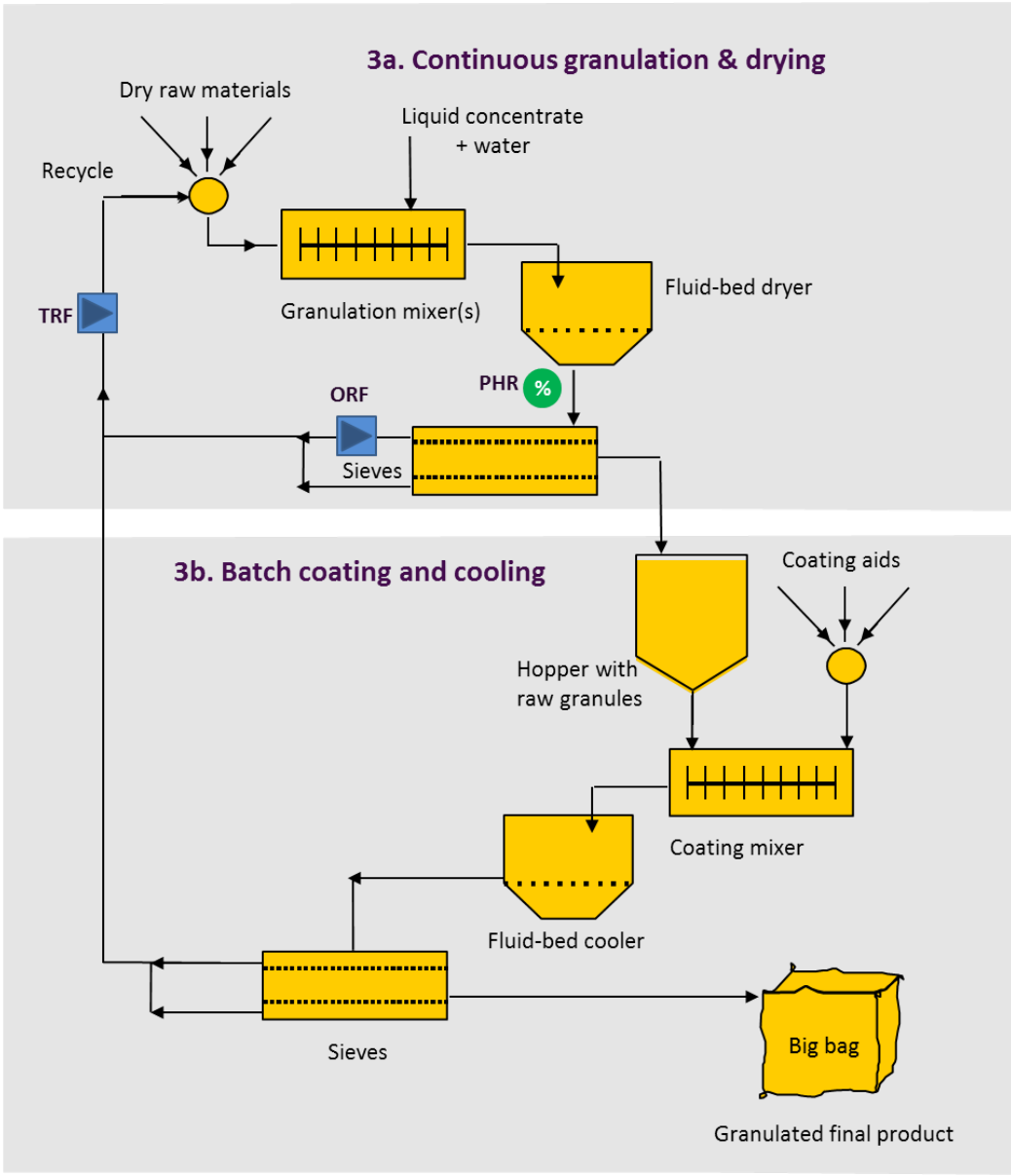


Figure 5-17 – Schematic overview of the manufacturing process of enzyme granulate with the approximate location of process tags: Total Recycle Flow (THR); Oversized Recycle Flow (ORF) and Powder Height Regulator (PHR).

The granulate product properties are primarily influenced by formulation properties and mixing within a granulation vessel [123]. The mechanism of granulation is very complex and challenging to model which makes the current operation very experienced-based [121]. The suspected reasons behind an uneven granulation

performance are numerous. For instance, inequality of the liquid concentrate introduces inconsistency into the granulation process. This means that the ability of the mixture to form a proper granulate can vary for each new batch of enzyme concentrate entering the process, as well as for subsequent portions of the same concentrate due to inhomogeneity within a tank volume. As a consequence no or poor granulate is formed or too large particles are found. This is later corrected by manipulation of the ratio between ingredients, but the abortive mixture increases the recycle pool which contributes to a capacity loss. Despite its widespread use, granulation processes are highly inefficient, and even modern industrial plants often operate with high recycle ratios. Thus, there is an economic incentive for a better understanding of granulation processes, leading to the more effective operation. New ways must be developed that allow for better supervision and to optimize the existing processes and improve monitoring systems.

5.2.1 Study of periodic patterns in the granulation processes

As a consequence of semi-continuous production system, there are no fixed or clear time-relations between successive unit operations in the granulation process flow. To get a better insight into system dynamics and move towards process optimization/variance reduction the serial correlation procedure has been used on historical datasets. Over fifty process tags were examined; however, the main attention was focused on the tag describing granulate fraction sieved out after a drying bed, not meeting the specification (both too coarse and too fine) represented by the tag 'Total Recycle Flow'. This parameter clearly indicates if the granulation is not performing well. Information that could be shared with regard to this study has been summarized in POSTER I. The percentage open of the rotary valve controlling the volume/amount of granulate dried in the fluid bed is given by the process tag 'Powder Height Regulator'. Location of all the tags discussed in the poster is indicated in Figure 5-17.

5.2.2 Lagging of the signals

The cross-correlation function (CCF) has been used in this project work with moderate success to identify delay times between unit operations involved in a (semi-)continuous granulation and drying process. The results strongly depend on the preceding filtering steps. It should be noted once more that part of the signal deformation happens during data compression which is irrecoverable and hard to

track back. Piecewise Linear De-trending (PLD, see Section 3.3.2) has been the method of choice for obtaining weakly stationary signals. However, the resulting autocorrelation and cross-correlation functions experienced broad peaks, and they were sometimes hard to read. It can be postulated that a more advanced pre-whitening of the signals could have been performed prior to CCF. However, if the seasonal component is removed from the time series the common stimulus which affects operations downstream to the drying bed is also removed. In such an approach, no significant cross-correlation between 'TRF' and 'PHR' is identified. Serial correlation analysis is certainly an interesting tool for examination of the delay times in different processes. However, application to the large-scale industrial process with numerous sources of variation, and differences in data density and quality requires more work than available in the framework of this Ph.D. project for this particular subject.

5.2.3 Summary

It was shown that an appropriate pre-processing of imperfect data can account for e.g. sensor inaccuracy, signal saturation or downtime. Different signal processing techniques, when applied wisely, can lead to interesting and otherwise undiscovered process knowledge. For instance, the examined 'TRF' signal showed some unexpected periodicity. Every 20 minutes the recycle flow is adjusted. This could hardly be registered without de-trending, marginally for most levels of pre-processing plus de-trending, but by far the clearest after smoothing plus de-trending.

It is obvious that good understanding of process fluctuations is very important. It allows efficient process management and production within narrow product specifications. Serial correlation analysis identified a previously unknown periodic adjustment in a fluidized bed unit operation. It was not possible to indisputably identify the root cause of the periodicity in a fluidized bed drying operation. The unknown stimulus is common for different process measurements related to the fluidized bed dryer. Hence, with the use of a cross-correlation function it was possible to estimate the lag time between the regulation of the powder height in the fluidized bed and its effect in the measurement of the total recycle flow being 5-6 minutes. Similarly, it was estimated that the two recycle flow measurements, total and oversized, are distanced by 1-2 minutes.

The reason for the observed periodic adjustment is most probably related to a mismatch between the location of the feedback signals and resulting delayed response in the control structures of the fill level in the bed. This thought is supported by the fact that no fluctuations in powder height and recycle are observed in the similar fluidized bed situated in another plant at Novozymes A/S, which is run under manual control of powder height.

6. Conclusions and Perspectives

This thesis focusses on solutions for a more extensive use of full-scale historical production data in mining, process optimization and problem-solving in the bioindustry. The research presented in this thesis has demonstrated that wise use of novel multivariate statistical tools can help to unlock the potential hidden in historical datasets.

The key focus of this thesis was a more optimal utilization of off-line analysis and not the development of on-line applications. Many bioindustrial processes are not yet fully understood. Frequently, it is not straightforward how the performance of the different operations should be measured and which factors provide higher efficiency and better product quality. Therefore, a sufficient degree of process understanding needs to be gained first, for instance with the methods proposed in this thesis work, before an on-line monitoring schemes can be implemented.

There is no master recipe that can tell you how this data mining challenge should be tackled. However, based on my experience in exploration of historical datasets from downstream processing of enzymes, I can recommend the following strategy:

1) Recognition and definition of the problem. Identification of:
   1.1) One entity (i.e. batch, run, campaign) and operation period (i.e. startup, steady-state, CIP)
   1.2) Variability and way(s) to quantify it
   1.3) Motivation/Goal
2) Data assembly and cleanup
3) Data validity check:
   3.1) Logging frequency to historian
   3.2) Raw data plotting
   3.3) Basic uni- and bi-variate statistics for the tags
   3.4) Identification of missing data
   3.5) Preliminary PCA on means matrix or VWU data

4) Model improvement
   4.1) Lagging/synchronization
   4.2) Removal of outliers
   4.3) Tag selection
   4.4) Signal filtering
5) Consider more advanced modeling approaches in the following order:
   5.1) Features extraction
   5.2) Multilevel approach
   5.3) Multiblock or multiway techniques
      (Zooming in on the smaller focus areas indicated by the preceding model)
6) Consult the outcome of the models with specialists via. proper visualization

Communicate results to the decision makers

Steps 2) and 3) can be automated to a considerable degree. However, steps involved in '4) Model improvement' and further down require supervision and careful selection of the appropriate approach. Many of the above steps are iterative, particularly in the beginning when a common understanding of the problem needs to be established between the data analyst and the process specialists. Rarely has the analyst sufficient process knowledge to progress on her own. Therefore, communication with the process experts along the way is essential. During such meetings, the chemometrician needs to visualize the digested data and the outcome of the models in an accessible way that everybody can relate to and interpret.

It would be a recommendation of this Ph.D. experience always to start with the basic LV methods before attempting more ambitious techniques. Simple PCA or PLS models can serve as a check on the data quality and our process understanding. They can help to spot clusters in the data, outliers and identify different abnormalities such as mistakes happening during data acquisition or periods of sensor failure. Based on this first investigation, one can e.g. decide which production periods and which tags should be excluded from the follow-up study. Potential limitations should also be identified at this stage. A primary objective of the historian archiving is to preserve the main operation while keeping the network traffic low. As a consequence, during the compression, data is frequently ironed-out from special occurrences and even possibly averaged beyond the point where alignment or synchronization techniques make any difference. Moreover, a substantial number of auxiliary process sensors

cannot be trusted either because they are broken or miscalibrated for extended periods of time. This issues, which are normally not critical for the ongoing process performance, handicaps the long-term data-driven optimization approaches. It is hence recommended to secure a proper surveillance system over the quality of the process data. For instance, an adequate data management should involve 1) a unified practices regarding the calibration of sensors, 2) ensuring that data is measured at the right place, 3) assuring that the DCS and historian compression settings are set correctly. Bearing in mind this sometimes poor quality and trustworthiness of archived production data, it is certainly better to use simple models which are more robust (less prone to overfitting) in a first investigation. Moreover, simple approaches can be exported to standard office software (such as Excel spreadsheets), and they are easier to explain to the non-chemometricians in an intuitive manner.

If the researcher and stakeholders are not satisfied with the directions given by the simple approaches, then it is the time to consider if somewhat more advanced techniques would offer a better explanation to the problem. Based on the preliminary examination, it can e.g. be decided to focus only on particular product variants or manufacturing periods in the following studies. At this stage in my UF flux study, we progressed by only focusing on three processing recipes (out of originally six) which were historically succeeding each other (PAPER II). The selected method should aim at handling the data in its natural structure. We found that it was advantageous to recognize the hierarchical arrangement in the examined datasets and to pursue a multilevel modeling. In this approach, level one is formed by process timestamps which are nested within the individual runs referred to as level two. In this method, we tracked the recipe-classification by labeling the data on the levels two and one according to the recipe in force at the time of production. Based on the outcome, we also extended the MSCA-P model to a third level, processing recipe. The advantage of this approach over the standard VWU-PCA is that the variation within the runs is not confounded with variation between the runs and likewise for between-run, and between-recipe levels. Parameters directly related to the flux decline have been found on the recipe and within levels, and the steep flux decline appears to be related to process temperature, and in particular to increased feed temperature.

Finally, in the work described in PAPER III, we looked closer into the group of production runs identified as the most problematic only, and handled it as a regression problem with UF flux as a response variable. Interpretability of the PLS model has been made more holistic and simplified by calculation of the lower and super-level multiblock parameters. As process variables were assigned to groups corresponding to distinct phases of the process or belonging to similar engineering type of sensors, it was considerably easier to study and interpret the behavior of these blocks rather than keeping track of individual loading values.

I have found this strategy of 'peeling off different layers' of the problem under study and zooming in on smaller sections of the process the right one. One may argue that the more data the better. However, from my industrial experience, there were too many sources of variability in the bigger datasets to come to any solid conclusion. On the reverse, by peeling of the problem, it was possible to see tendencies/patterns in the smaller groups of runs which were otherwise obscured. Moreover, it needs to be highlighted that a large number of variables is under closed-loop control in the production environment, which impose correlations. Therefore, a proper expert insight is crucial to decide which of the observed relations lead to new, unexpected findings and which are simply driven by the mechanics of the system. It is undoubtedly beneficial if the data analyst possesses the process insight or works in close collaboration with the experts.

My UF flux studies have pointed at the temperature of the feed and processing temperature on the recirculation loops as being predominantly responsible for the steep flux decline, but at the same time for the higher overall mean flux. Therefore, future work of the optimization engineers should be focused on finding a balance between the UF temperature and the parameters of individual UF orders (feed volume, the degree of concentration, etc.). Especially, the question should be addressed if, from an economic perspective, it is better to process the same order in several high throughput runs (involving extra cleaning) or in one extended run (without the need for extra CIP).

The natural hierarchy in process data is both a challenge and an opportunity. In PAPER II, we have used blocking in the row or time direction, whereas in PAPER III we applied blocking in the column or process tags direction. Both methods lead to decomposition of the complex data structures into intuitively interpretable solutions

by keeping the natural structure of the analyzed data. This might 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 analysis of the multivariate historical datasets even more. I am sure that such moderately advanced methods for multiset modeling will prove their worth and will become standard process chemometrics tools for exploratory problem solving and investigation.

In contrary to initial plans, I have not attempted to make a link between the process data collected during recovery and the subsequent granulation steps in the Ph.D. investigation. This is still a valid future perspective. However, to achieve this goal, Novozymes A/S needs to implement some more direct measure of the granulation performance. More research should be also be dedicated to proper alignment of data from the continuous granulation-drying process.

As stated in the beginning of this paragraph, it was not a primary ambition of this Ph.D. project to develop on-line applications of multivariate statistics. The closest to such an implementation was the NIR study for real-time monitoring of enzyme activity described in PAPER I. NIR spectroscopy is tailored for on-line applications, and it is one of the most commonly used spectrophotometric methods used in PAT developments. During monitoring, it can contribute by providing a faster feedback on process yield or product quality. Moreover, it often offers additional insight beyond its direct scope, leading to a better process understanding. It is however bound to the higher operational complexity of the implementations. NIR applications at Novozymes A/S have unfortunately been discontinued as it was not possible to obtain a satisfyingly stable and global solution for the related data management infrastructure and network.

I believe that companies which are better at turning data into information and decisions will dominate in the future. Biomanufacturing operations are still largely running on experience-based rather than science-based practices. PAT has the potential to change this. As bioindustry is regulated to a lower extent than e.g. the pharmaceutical industry, it should to the higher degree explore a variety of multivariate sensors and chemometrics applications in full-scale processing. The tightly controlled lab or pilot scale experiments rarely correspond to the conditions encountered in the large-scale where numerous sources of variation may occur. On

the large scale, the answer is PAT. Certainly, a significant amount of time and money needs to be allocated to qualify the personnel and to develop, implement and maintain the PAT solutions. However, in return, process understanding can be gained fast and 'myths' abolished when the data-driven frameworks are established and are running smoothly on the actual production records.

References

[1] T. Kourti, Process analysis and abnormal situation detection: from theory to practice, Control Systems, IEEE 22 (2002) 10-25.

[2] P. Kadlec, B. Gabrys, S. Strandt, Data-driven soft sensors in the process industry, Comput. Chem. Eng. 33 (2009) 795-814.

[3] A.K. Smilde, F.W.J. van den Berg, H.C.J. Hoefsloot, How to Choose the Right Process Analyzer, Anal. Chem. 74 (2002) 368 A-373 A.

[4] C.B. Lyndgaard, S.B. Engelsen, F.W.J. van den Berg, Real-time modeling of milk coagulation using in-line near infrared spectroscopy, J. Food Eng. 108 (2012) 345-352.

[5] J.H. Thygesen, F.W.J. van den Berg, Subspace methods for dynamic model estimation in PAT applications, J. Chemometrics 26 (2012) 435-441.

[6] J. Camacho, J. Picó, A. Ferrer, Multi-phase analysis framework for handling batch process data, J. Chemometrics 22 (2008) 632-643.

[7] W. Chew, P. Sharratt, Trends in process analytical technology, Anal. Methods 2 (2010) 1412-1438.

[8] B. Smith-Goettler, On-line PAT Applications of Spectroscopy in the Pharmaceutical Industry, Chapter 13 in: K.A. Bakeev (Ed.), Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries, 2nd ed., John Wiley & Sons, 2010, 439-461.

[9] T. Rajalahti, O.M. Kvalheim, Multivariate data analysis in pharmaceutics: a tutorial review, Int. J. Pharm. 417 (2011) 280-290.

[10] S. Charaniya, W.S. Hu, G. Karypis, Mining bioprocess data: opportunities and challenges, Trends Biotechnol. 26 (2008) 690-699.

[11] T. Kourti, Abnormal situation detection, three-way data and projection methods; robust data archiving and modeling for industrial applications, Annual Reviews in Control 27 (2003) 131-139.

[12] I. Miletic, S. Quinn, M. Dudzic, V. Vaculik, M. Champagne, An industrial perspective on implementing on-line applications of multivariate statistics, J. Process Control 14 (2004) 821-836.

[13] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, Int J Adapt Control Signal Process 19 (2005) 213-246.

[14] S.J. Doherty, A.J. Lange, Avoiding pitfalls with chemometrics and PAT in the pharmaceutical and biotech industries, TrAC Trends in Analytical Chemistry 25 (2006) 1097-1102.

[15] M. Champagne, M. Dudzic, Industrial use of multivariate statistical analysis for process monitoring and control, in Proc. Amer. Control Conf., Anchorage, AK, May 2002, vol.1 594-599.

[16] J.A. Lopes, J.C. Menezes, J.A. Westerhuis, A.K. Smilde, Multiblock PLS analysis of an industrial pharmaceutical process, Biotechnol. Bioeng. 80 (2002) 419-427.

[17] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemometrics Intellig. Lab. Syst. 28 (1995) 3-21.

[18] K.A. Bakeev, Future Trends in Process Analytical Chemistry, in: K.A. Bakeev (Ed.), Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries, 1st ed., Wiley-Blackwell, 2005, 424-444.

[19] OSIsoftLearning, Exception and compression full details (video file), Available at: http://www.youtube.com/watch?v=89hg2mme7S0 (Accessed 14 January 2016).

[20] J.F. MacGregor, T. Kourti, Statistical process control of multivariate processes, Control Eng. Pract. 3 (1995) 403-414.

[21] OSIsoft LCC, PI ProcessBook, PI DataLink, PI WebParts Course, 2010.

[22] T. Kourti, P. Nomikos, J.F. MacGregor, Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS, J. Process Control 5 (1995) 277-284.

[23] V. Gekas, Terminology for pressure-driven membrane operations, Desalination 68 (1988) 77-92.

[24] A. Eckner, A framework for the analysis of unevenly-spaced time series data, Available at: http://www.eckner.com/papers/unevenly_spaced_time_series_analysis.pdf (Accessed 29 November 2015).

[25] G. Zumbach, U. Müller, Operators on inhomogeneous time series, Int. J. Theor. Appl. Finance 4 (2001) 147-177.

[26] H. Madsen, Time Series Analysis, CRC Press, 2007.

[27] G.E. Box, W.G. Hunter, J.S. Hunter, Statistics for experimenters: An introduction to design, data analysis, and model building, Wiley & Sons, Inc.: New York, 1978.

[28] D.L. Massart, B.G. Vandeginste, L.M.C. Buydens, P.J. Lewi, J. Smeyers-Verbeke, Process Modelling and Sampling, Chapter 20 in: Handbook of Chemometrics and Qualimetrics: Part A, Elsevier Science Inc., 1997, pp. 587-642.

[29] D.M. Meko, Applied time series analysis, Course materials available at: www.ltrr.arizona.edu/~dmeko/geos585a.html (accessed 16 December 2015).

[30] T.C. Mills, The Foundations of Modern Time Series Analysis, Palgrave Macmillan UK, 2011.

[31] D.L. Massart, B.G. Vandeginste, L.M.C. Buydens, P.J. Lewi, J. Smeyers-Verbeke, Straight Line Regression and Calibration, Chapter 8 in: Handbook of Chemometrics and Qualimetrics: Part A, Elsevier Science Inc., 1998, pp. 171-230.

[32] T.J. Cleophas, A.H. Zwinderman, Time series, in: Statistics Applied to Clinical Studies, Springer, 2012, pp. 687-693.

[33] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden Day: San Francisco, 1976.

[34] D.L. Massart, B.G. Vandeginste, L.M.C. Buydens, P.J. Lewi, J. Smeyers-Verbeke, Control charts, Chapter 7 in: Handbook of Chemometrics and Qualimetrics: Part A, Elsevier Science Inc., 1998, pp. 151-170.

[35] P.A. Gorry, General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method, Anal. Chem. 62 (1990) 570-573.

[36] J. Trygg, Background Estimation, Denoising, and Preprocessing, Section 2.01 in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Volume 2, Elsevier, Oxford, 2009, pp. 1-8.

[37] P.H. Eilers, B.D. Marx, Flexible smoothing with B-splines and penalties, Statistical Science (1996) 89-102.

[38] P.H. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631-3636.

[39] F.W.J. van den Berg, Signal filter, MATLAB source code available at: http://www.models.life.ku.dk/~frans/Some_Matlab/Filter/ (Accessed 16 December, 2015).

[40] D.L. Massart, B.G. Vandeginste, L.M.C. Buydens, P.J. Lewi, J. Smeyers-Verbeke, Signal processing, Chapter 40 in: Handbook of Chemometrics and Qualimetrics: Part B, Elsevier Science Inc., 1998, pp. 507-574.

[41] D. Ronen, C.F. Sanders, H.S. Tan, P.R. Mort, F.J. Doyle III, Predictive dynamic modeling of key process variables in granulation processes using partial least squares approach, Ind Eng Chem Res 50 (2011) 1419-1426.

[42] C. Capilla, A. Ferrer, R. Romero, A. Hualda, Integration of statistical and engineering process control in a continuous polymerization process, Technometrics 41 (1999) 14-28.

[43] P. Nomikos, J.F. MacGregor, Monitoring batch processes using multiway principal component analysis, AIChE J. 40 (1994) 1361-1375.

[44] P. Nomikos, J.F. MacGregor, Multi-way partial least squares in monitoring batch processes, Chemometrics Intellig. Lab. Syst. 30 (1995) 97-108.

[45] T. Kourti, Process analytical technology beyond real-time analyzers: The role of multivariate analysis, Crit. Rev. Anal. Chem. 36 (2006) 257-278.

[46] J. Camacho, J. Pico, A. Ferrer, Bilinear modelling of batch processes. Part I: theoretical discussion, J. Chemometrics 22 (2008) 299-308.

[47] J.M. González-Martínez, R. Vitale, O.E. de Noord, A. Ferrer, Effect of synchronization on bilinear batch process modeling, Ind Eng Chem Res 53 (2014) 4339-4351.

[48] J.M. González-Martínez, O.E. de Noord, A. Ferrer, Multisynchro: a novel approach for batch synchronization in scenarios of multiple asynchronisms, J. Chemometrics 28 (2014) 462-475.

[49] C. Duchesne, T. Kourti, J.F. MacGregor, Multivariate SPC for startups and grade transitions, AIChE J. 48 (2002) 2890-2901.

[50] Y. Zhang, M. Dudzic, V. Vaculik, Integrated monitoring solution to start-up and run-time operations for continuous casting, Annual Reviews in Control 27 (2003) 141-149.

[51] T. Kourti, Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions, J. Chemometrics 17 (2003) 93-109.

[52] S. Wold, N. Kettaneh-Wold, J.F. MacGregor, K.G. Dunn, Batch Process Modeling and MSPC, Section 2.10 in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 163-197.

[53] S.W. Andersen, G.C. Runger, Automated feature extraction from profiles with application to a batch fermentation process, Journal of the Royal Statistical Society: Series C (Applied Statistics) 61 (2012) 327-344.

[54] C. Undey, B.A. Williams, A. Cinar, Monitoring of batch pharmaceutical fermentations: Data synchronization, landmark alignment, and real-time monitoring, in: Proc. 15th IFAC World Congress, Barcelona, Spain, 2002.

[55] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Statistical Description of Data, Chapter 14 in: Numerical Recipes in C, Cambridge university press Cambridge, 1996, pp. 609-655.

[56] R.H. Jellema, Variable Shift and Alignment, Section 2.06 in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier, Oxford, 2009, pp 85.-108.

[57] R.A. Davis, A.J. Charlton, J. Godward, S.A. Jones, M. Harrison, J.C. Wilson, Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform, Chemometrics Intellig. Lab. Syst. 85 (2007) 144-154.

[58] A. Bogomolov, Multivariate process trajectories: capture, resolution and analysis, Chemometrics Intellig. Lab. Syst. 108 (2011) 49-63.

[59] E.M. Qannari, I. Wakeling, P. Courcoux, H.J. MacFie, Defining the underlying sensory dimensions, Food Quality and Preference 11 (2000) 151-154.

[60] B.M. Wise, N.B. Gallagher, The process chemometrics approach to process monitoring and fault detection, J. Process Control 6 (1996) 329-348.

[61] K. Pearson, Principal components analysis, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6 (1901) 559.

[62] H. Hotelling, Analysis of a complex of statistical variables into principal components., J. Educ. Psychol. 24 (1933) 417.

[63] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics Intellig. Lab. Syst. 2 (1987) 37-52.

[64] R. Bro, A.K. Smilde, Principal component analysis, Analytical Methods 6 (2014) 2812-2831.

[65] J.V. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, The Canadian Journal of Chemical Engineering 69 (1991) 35-47.

[66] T. Kourti, J.F. MacGregor, Multivariate SPC methods for process and product monitoring, Journal of Quality Technology 28 (1996) 409-428.

[67] L. Zullo, Validation and verification of continuous plants operating modes using multivariate statistical methods, Comput. Chem. Eng. 20 (1996) 683-688.

[68] A. Raich, A. Cinar, Statistical process monitoring and disturbance diagnosis in multivariable continuous processes, AIChE J. 42 (1996) 995-1009.

[69] A.J. Burnham, R. Viveros, J.F. MacGregor, Frameworks for latent variable multivariate regression, J. Chemometrics 10 (1996) 31-45.

[70] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics Intellig. Lab. Syst. 58 (2001) 109-130.

[71] R. Bro, A.K. Smilde, Centering and scaling in component analysis, J. Chemometrics 17 (2003) 16-33.

[72] H.F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (1958) 187-200.

[73] J. Valsiner, P.C. Molenaar, M. Lyra, N. Chaudhary, Dynamic Process Methodology in the Social and Developmental Sciences, Springer, 2009.

[74] M.E. Timmerman, Multilevel component analysis, Br. J. Math. Stat. Psychol. 59 (2006) 301-320.

[75] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometrics 12 (1998) 301-321.

[76] J.A. Westerhuis, A.K. Smilde, Deflation in multiblock PLS, J. Chemometrics 15 (2001) 485-493.

[77] S.J. Qin, S. Valle, M.J. Piovoso, On unifying multiblock analysis with application to decentralized process monitoring, J. Chemometrics 15 (2001) 715-742.

[78] O.E. de Noord, E.H. Theobald, Multilevel component analysis and multilevel PLS of chemical process data, J. Chemometrics 19 (2005) 301-307.

[79] J.F. MacGregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, Process monitoring and diagnosis by multiblock PLS methods, AIChE J. 40 (1994) 826-838.

[80] J.J. Jansen, H.C. Hoefsloot, J. van der Greef, M.E. Timmerman, A.K. Smilde, Multilevel component analysis of time-resolved metabolic fingerprinting data, Anal. Chim. Acta 530 (2005) 173-183.

[81] C.M. Andersen, R. Bro, Variable selection in regression—a tutorial, J. Chemometrics 24 (2010) 728-737.

[82] I. Chong, C. Jun, Performance of some variable selection methods when multicollinearity is present, Chemometrics Intellig. Lab. Syst. 78 (2005) 103-112.

[83] T. Rajalahti, R. Arneberg, F.S. Berven, K. Myhr, R.J. Ulvik, O.M. Kvalheim, Biomarker discovery in mass spectral profiles by means of selectivity ratio plot, Chemometrics Intellig. Lab. Syst. 95 (2009) 35-48.

[84] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S. Engelsen, Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413-419.

[85] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for monitoring batch processes, Technometrics 37 (1995) 41-59.

[86] J. Camacho, Tutorial: Multi-Phase Framework Toolbox, MP-Toolbox, Version 1. 0 (2012).

[87] J. Chen, K. Liu, On-line batch process monitoring using dynamic PCA and dynamic PLS models, Chemical Engineering Science 57 (2002) 63-75.

[88] W. Ku, R.H. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, Chemometrics Intellig. Lab. Syst. 30 (1995) 179-196.

[89] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, J. Chemometrics 17 (2003) 323-337.

[90] T. Kourti, Multivariate Statistical Process Control and Process Control, Using Latent Variables, Section 4.02 in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford 2009, pp. 54-126.

[91] E. Martin, A. Morris, Non-parametric confidence bounds for process performance monitoring charts, J. Process Control 6 (1996) 349-358.

[92] A. Ferrer-Riquelme, Statistical Control of Measures and Processes, Section 1.04 in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Oxford, 2009, pp. 97-126.

[93] D.C. Montgomery, Introduction to Statistical Quality Control, John Wiley & Sons, 2007.

[94] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Standardized Q-statistic for improved sensitivity in the monitoring of residuals in MSPC, J. Chemometrics 14 (2000) 335-349.

[95] S. Bersimis, S. Psarakis, J. Panaretos, Multivariate statistical process control charts: an overview, Qual. Reliab. Eng. Int. 23 (2007) 517-543.

[96] B.M. Wise, R.T. Roginski, A Calibration Model Maintenance Roadmap, IFAC-Papers On Line 48 (2015) 260-265.

[97] K. Aunstrup, Vejen Til Novozymes, ISBN 87-988324-1-7 ed., Novozymes A/S, 2001.

[98] J. Rogers, E.L.H. Petersen, Enzymes bring the magic of nature into detergents, Inform-International News on Fats Oils and Related Materials 16 (2005) 324-325.

[99] P.P. Mortensen, R. Bro, Real-time monitoring and chemical profiling of a cultivation process, Chemometrics Intellig. Lab. Syst. 84 (2006) 106-113.

[100] A.E. Cervera-Padrell, N. Petersen, A.E. Lantz, A. Larsen, K.V. Gernaey, Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation, Biotechnol. Prog. 25 (2009) 1561-1581.

[101] L.R. Formenti, A. Nørregaard, A. Bolic, D.Q. Hernandez, T. Hagemann, A. Heins, H. Larsson, L. Mears, M. Mauricio-Iglesias, U. Krühne, Challenges in industrial fermentation technology research, Biotechnology Journal 9 (2014) 727-738.

[102] K.V. Gernaey, A.E. Cervera-Padrell, J.M. Woodley, A perspective on PSE in pharmaceutical process development and innovation, Comput. Chem. Eng. 42 (2012) 15-29.

[103] P.P. Mortensen, Process Analytical Chemistry (PAC)-Opportunities and Problems for Bio-industrial Implementation, Ph.D. Thesis, Esbjerg: Aalborg University (2006).

[104] S. Bondesson, The effect of the flocculation chemicals on the downstream UF performance, Lund University (2013), Manuscript available online at: http://www.chemeng.lth.se/exjobb/E688.pdf (Accessed 14 January 2016).

[105] M. Cheryan, Ultrafiltration and Microfiltration Handbook, Technomic Pub. Co, Inc., Lancaster, 1998, pp. 264.

[106] C. Charcosset, Membrane processes in biotechnology: an overview, Biotechnol. Adv. 24 (2006) 482-492.

[107] J. Wagner, Membrane Filtration Handbook: Practical Tips and Hints, Osmonics Minnetonka, MN, 2001.

[108] G. Bylund, Dairy Processing Handbook, Tetra Pak Processing Systems AB, Lund, Sweden, 1995.

[109] J. Brown, A. Gilby, B. Lipták, T. Cardis, E. Baughman, Infrared and Near-Infrared Analyzers, in: B.G. Lipták (Ed.), Instrument Engineers' Handbook: Process Measurements and Analysis, 4th ed., 2003, pp. 1369-1387.

[110] E. Baughman, Process analytical chemistry: introduction and historical perspective, Chapter 1 in: K.A. Bakeev (Ed.), Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries, 1st ed., Wiley-Blackwell, 2005, pp. 1-11.

[111] T.H.A. Berg, J.C. Knudsen, R. Ipsen, F.W.J. van den Berg, H.H. Holst, A. Tolkach, Investigation of consecutive fouling and cleaning cycles of ultrafiltration membranes used for whey processing, International Journal of Food Engineering 10 (2014) 367-381.

[112] J.K. Jensen, J.M.A. Rubio, S.B. Engelsen, van den Berg, F.W.J., Protein residual fouling identification on UF membranes using ATR-FT-IR and multivariate curve resolution, Chemometrics Intellig. Lab. Syst. 144 (2015) 39-47.

[113] G. van den Berg, C. Smolders, Flux decline in ultrafiltration processes, Desalination 77 (1990) 101-133.

[114] J. Linkhorst, W.J. Lewis, Workshop on membrane fouling and monitoring: a summary, Desalination and Water Treatment 51 (2013) 6401-6406.

[115] A. Marshall, P. Munro, G. Trägårdh, The effect of protein fouling in microfiltration and ultrafiltration on permeate flux, protein retention and selectivity: a literature review, Desalination 91 (1993) 65-108.

[116] W. Guo, H. Ngo, J. Li, A mini-review on membrane fouling, Bioresour. Technol. 122 (2012) 27-34.

[117] A.T. Fane, R. Wang, Y. Jia, Membrane Technology: Past, Present and Future, in: Membrane Technology: Past, Present and Future Membrane and Desalination Technologies, Springer, 2011, pp. 1-45.

[118] W.L. McCabe, J.C. Smith, P. Harriott, Unit Operations of Chemical Engineering, McGraw-Hill Education, 2005.

[119] P. Bacchin, P. Aimar, R.W. Field, Critical and sustainable fluxes: theory, experiments and applications, J. Membr. Sci. 281 (2006) 42-69.

[120] R.B. Cattell, The scree test for the number of factors, Multivariate Behavioral Research 1 (1966) 245-276.

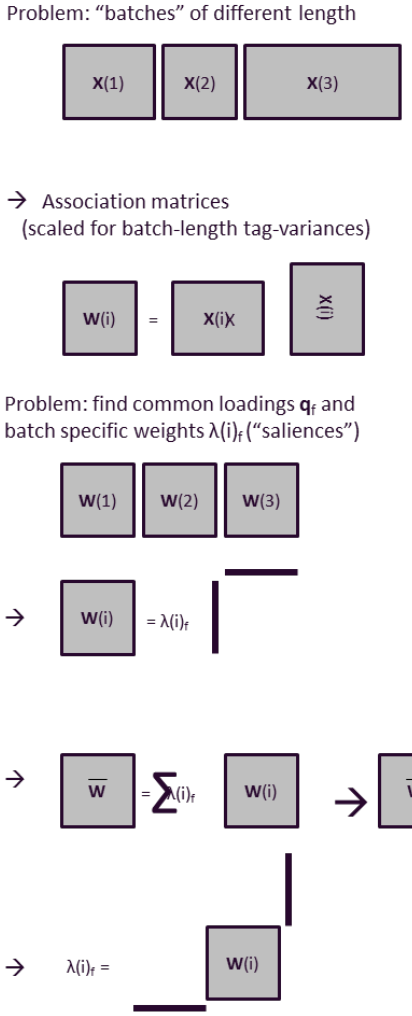
[121] O. Kirk, T. Damhus, T.V. Borchert, C.C. Fuglsang, H.S. Olsen, T.T. Hansen, H. Lund, H.E. Schiff, L.K. Nielsen, Enzyme Applications, Industrial, in: Enzyme Applications, Industrial Kirk-Othmer Encyclopedia of Chemical Technology, John Wiley & Sons, Inc., 2000.

[122] J. Litster, B. Ennis, The Science and Engineering of Granulation Processes, Springer Science & Business Media, 2013.

[123] D.A. Pohlman, J.D. Litster, Coalescence model for induction growth behavior in high shear granulation, Powder Technol 270, Part B (2015) 435-444.

Appendix

Appendix 1 - Common Components and Specific Weights Analysis



$$\boldsymbol{X}(i)\boldsymbol{X}(i)^T = \boldsymbol{W}(i) = \boldsymbol{Q}\boldsymbol{\Lambda}(i)\boldsymbol{Q}^T = \sum_{f=1}^{F} \boldsymbol{q}_f$$

$$\rightarrow L = \sum_{i=1}^{I} \left\| \boldsymbol{W}(i) - \sum_{f=1}^{F} \boldsymbol{q}_f \lambda(i)_f \boldsymbol{q}_f^T \right\|^2$$

$$\lambda(i)_f = \boldsymbol{rand}$$

$$\overline{\boldsymbol{W}} = \sum_{i=1}^{I} \lambda(i)_f \boldsymbol{W}(i)$$

$$\boldsymbol{q}_f = eig(\overline{\boldsymbol{W}})$$

$$\lambda(i)_f = \boldsymbol{q}_f^T \boldsymbol{W}(i) \boldsymbol{q}_f$$

$$\boldsymbol{X}(i) = \boldsymbol{X}(i) - \boldsymbol{q}_f \boldsymbol{q}_f^T \boldsymbol{X}(i)$$
$$\rightarrow \boldsymbol{W}(i) = \boldsymbol{X}(i)\boldsymbol{X}(i)^T$$
$$\rightarrow f = f + 1$$

Figure 1A – Algorithm and scheme of the working manner of the CCSWA.

PAPER I

A. Klimkiewicz, P.P. Mortensen, C.B. Zachariassen, F.W.J. van den Berg

Monitoring an enzyme purification process using on-line and in-line NIR measurements

Chemometrics and Intelligent Laboratory Systems, 132 (2014), 30–38

Chemometrics and Intelligent Laboratory Systems 132 (2014) 30–38

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Monitoring an enzyme purification process using on-line and in-line NIR measurements

Anna Klimkiewicz [a,b,*], Peter Paasch Mortensen [a], Christian B. Zachariassen [a], Frans W.J. van den Berg [b]

[a] Maintenance & Technology, Novozymes A/S, Hallas Alle, DK-4400 Kalundborg, Denmark
[b] Spectroscopy & Chemometrics, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 22 October 2013
Received in revised form 17 December 2013
Accepted 6 January 2014
Available online 16 January 2014

Keywords:
On-line NIR
In-line NIR
Enzyme activity
Ultrafiltration
Process monitoring
Real-time

ABSTRACT

The key parameter during purification of enzymes is the strength of the intermediates and products. Employing near infrared spectroscopy, enzyme activity of the concentrate can be obtained much faster than by traditional off-line analysis in a central laboratory. This paper describes the development of a monitoring system for the enzyme activity in a recovery plant and provides a comparison between two *real-time* probe setups (on-line v. in-line) in a full scale application. The focus in the paper will be on the chemometric calibration development plus validation and the application of the models for monitoring purposes. For this investigation, four different types of industrial enzymes have been sampled over a period of ten months. Real-time output of the partial least squares regression models was used along with conventional process data generated to evaluate the pros and cons of the in-line and on-line setup. Both implementations deliver good results in monitoring the ultrafiltration process, but problems (precipitation/phase transition) were occasionally encountered for the on-line arrangement.

1. Introduction

Industrial enzyme production is a complex discipline where numerous critical factors have to be controlled within narrow specifications in order to ensure a profitable outcome. The enzyme manufacturing process can be described by three core producing and formulation steps: 1. Cultivation of enzyme producing organisms, 2. Recovery of enzymes from the culture broth, 3. Formulation of enzymes into end-products. In most production sites, these three steps are almost *autonomous factories* or operations, but there obviously is a strong effect of upstream process performance on downstream execution, i.e. variations in the enzyme activity of the culture broth or the presence of foreign compounds having an adverse effect during the purification steps. When the cultivation is completed, the mixture of cells, nutrients and enzymes is subjected to downstream processing where enzymes are separated from the broth, condensed, purified and stabilized. The factory involving this sequence of unit operations, resulting in the concentrated enzyme product, is traditionally called *recovery*. The liquid enzyme product downstream from this processing stage can undergo formulation to the granulated dry product or be sold as a finished liquid product.

In recovery, final product specifications such as enzyme activity, color or turbidity are of high importance. In particular, the yield balances over different separation steps become more and more critical from an economic perspective, as the strength of the intermediary product increases moving downstream. With this in mind, relatively expensive process near infrared (NIR) instruments are considered more and more for supervision of quality and efficiency of the industrial operations. In the Novozymes production facilities at Kalundborg, Denmark, a project was initiated which investigates the potential of NIR technology for monitoring the enzyme activity in various liquid streams involved in recovery. Employing NIR spectroscopy, the indirect or inferential parameter enzyme activity in the concentrates can be obtained much faster than by traditional off-line analysis in a central laboratory. The work presented here focuses on the process stream coming from ultrafiltration (UF) concentrate and evaluates (a) which of two *real-time* NIR flow cell setups is the preferable implementation and (b) if statistical process monitoring can be used for early warning/fault detection.

Even though NIR is not normally considered very selective, due to, e.g. overlapping bands and a strong water signal, it has been shown numerous times to be a powerful quantitative tool. This is mainly attributed to the high signal-to-noise ratio. NIR spectroscopy is also tailored for timely measurements in dynamic systems as spectra can be acquired fast, without sample preparation, in non-invasive modes [1–3]. In addition, the light in the NIR region holds the huge advantage of transmission over larger distances by optical fibers [4], incorporating the cost–advantage of multiplexing. Another major benefit of optical fibers is elimination of complicated sampling systems to bring specimens from the process to the spectrometer. Indeed, it is possible to locate the equipment remotely in a safe environment using compact sample cells inside existing process streams (a *window* into the system). A serious disadvantage of NIR is that construction of a good multivariate

* Corresponding author at: Maintenance & Technology, Novozymes A/S, Hallas Alle, DK-4400 Kalundborg, Denmark. Tel.: +45 30771436.
E-mail address: akcz@novozymes.com (A. Klimkiewicz).

0169-7439/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2014.01.002

calibration model requires a considerable amount of effort, time and thus money.

Initial studies have shown that there is a strong relation between NIR spectra and enzyme activity in the UF concentrate. It was possible to develop satisfying calibration models for four examined enzyme products. Nevertheless, since large sample sets generally increase the robustness of calibration, development of a global model was pursued in this investigation. Standardizing several quality control (QC) parameters for enzyme activity into one global QC parameter proved successful, and these global models will be used in this comparative study.

Flow sample cells were directly mounted in the process lines (called *in-line* analysis) and introduced via a fast loop with sample conditioning system (*on-line* analysis). The challenge is to provide a timely and representative analysis result of the enzyme activity in the concentrate stream resulting from UF and to detect (predict) undesired behavior of the system as early as possible. The location of the sample point should thus be such that the analyzed stream represents the overall condition of the process to be monitored (or controlled) [5]. The local effects, such as orientation of the pipeline, the physical location of the sample probe and the design of the sample cell, should be used to favor the representativeness of the process measurements. The flow cells and sampling taps of the setups examined in this research were arranged to meet the representativeness and timeliness conditions [6,7].

2. Material and methods

2.1. NIR instrumentation

The analyzer system consists of an FTPA2000-260 Fourier Transform NIR spectrophotometer (ABB Bomem, Quebec, Canada) equipped with an InGaAs detector situated in a temperature controlled environment (thermoelectrically cooled detector box). The operational range of the instrument is 5300–10,000 cm^{-1}. This multiplexer setup allows for eight NIR channels, visited sequentially. One channel is used for internal control of the spectrometer leaving seven channels for monitoring inside the recovery processes. The light from the instrument was transported to and from the NIR transmission flow cells via approximately 30 m of 500 μm core diameter single strand fiber optic cables. The sampling frequency was not fixed/equidistance as a consequence of the multiplexing procedure that was used.

2.2. Process measurements

Fig. 1 presents the position of the two NIR probe arrangements in relation to the UF unit plus some related conventional process measurements of importance. These process signals are in this manuscript only consulted to account for abnormalities in the UF operation or peculiarities in NIR spectra and/or predictions, not for model building.

Both NIR flow cells – adjusted to different dimensions of the surrounding pipelines, have the exact same design and optical path length – are situated just downstream of the UF outlet (Fig. 1). The in-line flow cell is mounted directly in the process flow coming from a heat exchanger, whereas the on-line flow cell is installed in a stop-flow sampling fast-loop, bypassing the heat exchanger. The flow cells are mounted in horizontal tubing, where ideally a vertically oriented upwards flow should have been selected [7]. This arrangement was however unpractical with the existing unit design and process layout, but the high flow of UF concentrate relative to the small diameters of the pipelines ensures a homogeneous composition across the process stream. Manually operated sampling taps are situated directly after each flow cell to withdraw a specimen simultaneously with the recording of spectra. A programmable logic controller (PLC) signals the NIR instrumentation to perform a measurement from the different flow cells. The PLC also controls the temperature adjustment of the material in

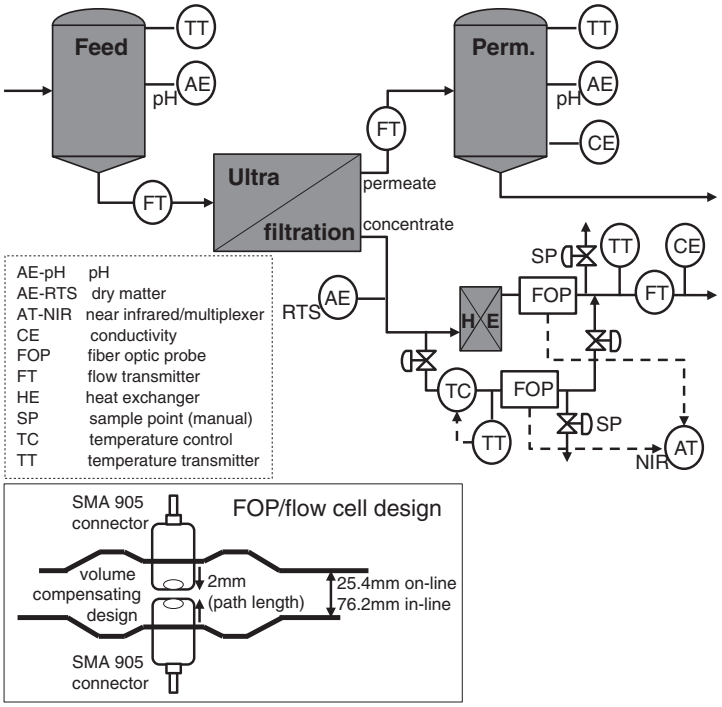


Fig. 1. Location of NIR flow cells in relation to the ultrafiltration unit.

the on-line (flow cell) bypass after closing two valves (sample temperature is elevated to just above maximal expected process stream temperature).

2.3. Data collection

Four different types of intermediate enzyme product have been sampled over a period of ten months—a total of sixty-two production batches varying in enzyme protein concentration (Table 1). Thirty-seven (in-line) and twenty-nine (on-line) of these batches were utilized in calibration model building; eighteen and sixteen production campaigns were used for test set validation (all test batch runs are at the end of the ten month sampling period). Data pairs of NIR spectra and reference samples (separate specimen for in-line and on-line; analyzed in the central laboratory; process time logging is used for alignment between NIR and manual sample extraction) are used for modeling. Process data from conventional measurements (Fig. 1) for fifty-two production batches were examined for monitoring of the UF process with the sampling frequency/collection times dictated by the NIR instrument/PLC.

All reference analyses performed were enzyme-specific assays, which estimate the activity in a given sample by the use of a substrate for which the particular enzyme has affinity. The enzyme releases, e.g. a color component from the substrate, which is determined by spectroscopic analysis. This number was standardized against the enzyme activity of the pure enzyme resulting in a value for enzyme protein expressed as %(w/w). The percentage of total enzyme protein in a specimen is derived from the enzyme activity from that sample via the following equation:

$$\text{Total Enzyme Protein } (\%) = f_{Rescale}\left(\frac{\text{Enzyme Activity}\left(\frac{Unit}{g\ sample}\right)}{\text{Specific Activity}\left(\frac{Unit}{g\ pure\ protein}\right)} \times 100\%\right)$$

It should be noted that the Total Enzyme Protein values are rescaled (further details are withheld for proprietary reasons) and should thus be thought of as having the arbitrary unit *percentage*.

2.4. Data analysis

NIR data is used for construction of partial least squares regression (PLS) calibration models for predicting the total enzyme protein (*global models*). All spectra are smoothed with so-called Savitzky–Golay filters (SG; window size 13, 2nd-order polynomial fitting) while calculating the first derivative and mean centered. This spectral preprocessing was selected after careful examination of different alternatives [8] and generally led to less complex models. PLS models were generated and validated using Matlab (version 8.0.0.783 (R2012b), Mathworks, USA) and the PLS Toolbox (Version 7.3.1, Eigenvector Research Inc., Manson, WA, USA) plus the iPLS algorithm for variable/interval selection [9]. All models are cross validated during calibration development to determine the appropriate model complexity, leaving out one batch at a time. Prediction residuals (cross validation or test set) were calculated as the difference between the reconstructed data and the original laboratory references. They represent a measure of how well an existing model fits to the new (future) samples.

Table 1
Data sets used in calibration and validation of the models.

Product	A[a]	B[b]	C[a]	D[b]
Plotting symbol	Triangle (red)	Square (blue)	Dot (green)	Lozenge (black)
Calibration[d]				
In-line[e]	11 (3)[c]	17	6 (1)[c]	7
On-line	11 (2)[c]	12 (2)[c]	5 (1)[c]	6
Validation[f]				
In-line[e]	3	7	4 (1)[c]	6 (1)[c]
On-line	3	6	3	6 (2)[c]

[a] amylase category, concentration range 1–5%.
[b] protease category, concentration range 10–20%.
[c] number of outliers removed during modeling, see Section 3.1.
[d] representing 42 production runs.
[e] fewer on-line samples due to missing reference analysis.
[f] representing 20 production runs.

3. Results and discussion

3.1. Model building

Spectra registered by the two flow cells and used for calibration of on-line and in-line PLS models are presented in Fig. 2. All except one of the excluded samples were spectral outliers exhibiting a high baseline absorbance and a saturation effect in the water peak region (spectra not included in Fig. 2). One outlier was removed due to a high prediction residual. Discarded outliers were later investigated with diagnostic plots. It should be noted that some of the regions in the spectra were excluded before variable selection. This includes part of the water peak region due to persistent absorbance saturation observed on the detector response and the region above 9000 cm^{-1}. At this stage, despite the removal of high absorbance outliers, there are still cases with a high baseline absorbance for the in-line probe. On the other hand, seven of the on-line samples were lost due to imperfections of the (manual) sampling setup. Hence, based on inspection of the calibration spectra, it cannot be concluded beforehand that either of the two flow cell positions is more prone to scattering due to, e.g. entrapment of air bubbles between the probe heads (Fig. 1), fouling or phase transition. Generally, baseline offsets and slopes are effectively removed by the SG preprocessing (Fig. 3).

Those parts of the spectra which were judged irrelevant for prediction were discarded using the iPLS algorithm. The selection method was applied separately for on-line and in-line calibration. A window of five adjacent variables has been used which corresponds to a wavenumber window width of 62 cm^{-1} (or 21 nm). The iPLS algorithm was run in the *reverse* mode meaning that intervals were successively removed from the analysis [9]. The interval, which when left out resulted in a remaining dataset producing a model with the smallest RMSECV, is selected for permanent exclusion. Successive cycles remove the next interval, etc. The average spectra after preprocessing for each of the four

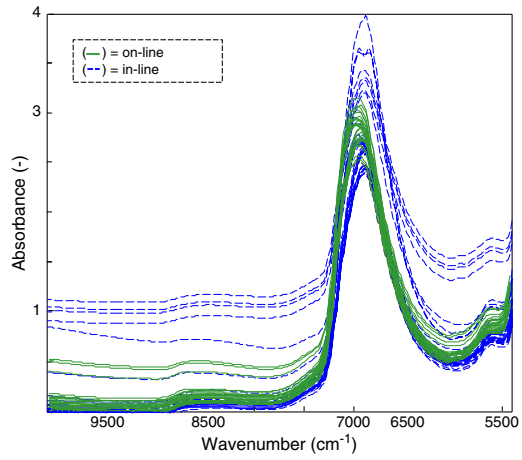


Fig. 2. Comparison of in-line and on-line spectra.

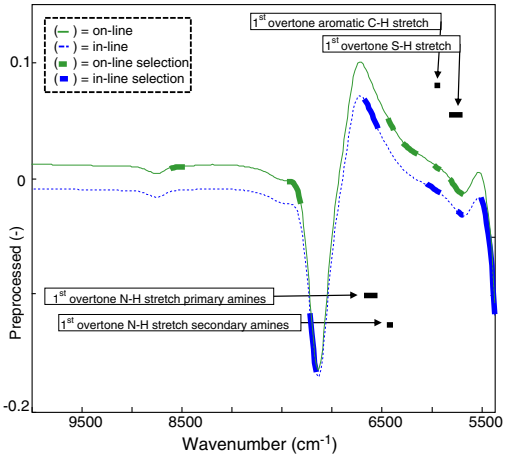


Fig. 3. Selected/representative spectra for in-line and on-line after preprocessing (plotted with offsets for clarity). Variables (wavenumbers) selected by the iPLS algorithm are indicated as bold line-segments and tentative spectral interpretation reported.

examined product types differed mainly in the intensity of the water band regions (results not shown). The spectra for proteases (products B and D) and amylases (products A and C) are different in the water peak region, explainable by the fact that the enzyme process streams of amylases are considerably more diluted than for proteases. Proteases exhibit a small *shoulder-band* found at 5975 cm^{-1} (5930–5920 cm^{-1} in the case of the un-processed spectra). This appears to be related to the first overtone of the aromatic C–H stretch (assigned to 1685 nm, [1]). Both of the final models use the information in that specific region. The second region taken into account by both models involves spectral variations between 5832 and 5694 cm^{-1}. It can be related to the reported position of a first overtone S–H stretch vibration at 1740 nm (~5748 cm^{-1}, [1]), a functional group present in active centers of various enzymes. Overtones related to protein structure, such as the first overtone of N–H stretch of primary and secondary amines, can account for selection of regions 6681–6542 cm^{-1} in the in-line and 6449–6388 cm^{-1} in the on-line model. The effects of removal of the outliers and the subsequent variable selection on model performance are summarized in Table 2.

The final in-line PLS calibration model after variable selection uses forty absorbance values and is of lower complexity with respect to the number of latent variables (LV) while the root mean square error of cross-validation (RMSECV) is also improved considerably from 0.88% to 0.64%. The on-line calibration uses fifty absorbance values with a prediction improvement from an RMSECV of 0.77% to 0.62%.

The discrepancy between the numbers of variables and components required in the models may be explained by the fact that the in-line calibration needs to account for drastically different temperatures (span approximately 30 °C for this product range), while the on-line setup was calibrated for a fixed, elevated temperature. Consequently, model complexity is reduced considerably (from 6 to 3 LVs) by variable selection, and one additional component may be required to account for the temperature deviations encountered in-line. In liquids containing water as a major component (as it is the case with enzyme concentrate), the effect of temperature changes is particularly evident as it affects the degree or strength of hydrogen bonding and the hydration status of all constituents, and these changes influence the wavelengths at which overtones or combination tones appear [10]. However, the use of selected spectral regions together with global spectral de-trending (the first derivatives) tends to make the analyte signals more dominant compared to the inherent effect of temperature change.

Prediction performance of the two final models is presented in Fig. 4. The obtained results were deemed satisfactory, taking into consideration uncertainties in laboratory reference analysis and the ability to predict the content of total enzyme protein in four different products. The gap in the calibration range (from approximately 5 to 10%) does not seem to have a negative impact on the performance.

3.2. Model testing

Model performance over an extended period of time has been investigated to check for possible instrumental drift or lack of robustness towards batch-to-batch variability. Stability of instruments can change over time, and small continuous changes (e.g. instrumental drift due to light source weakening) or sudden jolts (e.g. response shifts caused by repairs/replacements, or unexplained transformations in the process surrounding the spectrometer or probes) can cause the signal of an instrument to alter and affect the prediction errors. The (long-term) robustness of calibrations – here based on batches produced after the calibration period – is therefore a subject of high relevance in process chemometrics [11]. This was investigated using a test set not included in the calibration step, and the outcome is presented in Fig. 5. No deterioration of model performance with time was observed. The in-line

Table 2
Model development and performance.

	In-line flow cell[a]			On-line flow cell[a]		
	Model I	Model II[b]	Model III[c]	Model I	Model II[b]	Model III[c]
Spectral range (cm^{-1})	5385–6681 7113–8995	5385–6681 7113–8995	5385–5524 5693–5755 5925–6064 6542–6681 7159–7221	5385–6650 7128–8995	5385–6650 7128–8995	5693–5832 5925–5987 6156–6295 6388–6449 7314–7452 8471–8610
Variables	208	208	40	205	205	50
Calibration samples	41	37	37	34	29	29
Validation samples	20	18	18	18	16	16
LVs	6	4	3	2	2	2
RMSEC	0.62	0.62	0.53	0.91	0.68	0.54
RMSECV	2.10	0.88	0.64	1.24	0.77	0.62
RMSEP	1.09	0.91	0.83	2.27	1.48	0.98
R^2 CV	0.91	0.98	0.99	0.96	0.98	0.99
R^2 pred.	0.97	0.98	0.98	0.88	0.94	0.97

a) all models based on first derivative data with mean centering; b) samples with high baseline removed; c) samples with high baseline removed and reverse iPLS used for variable selection, final model.

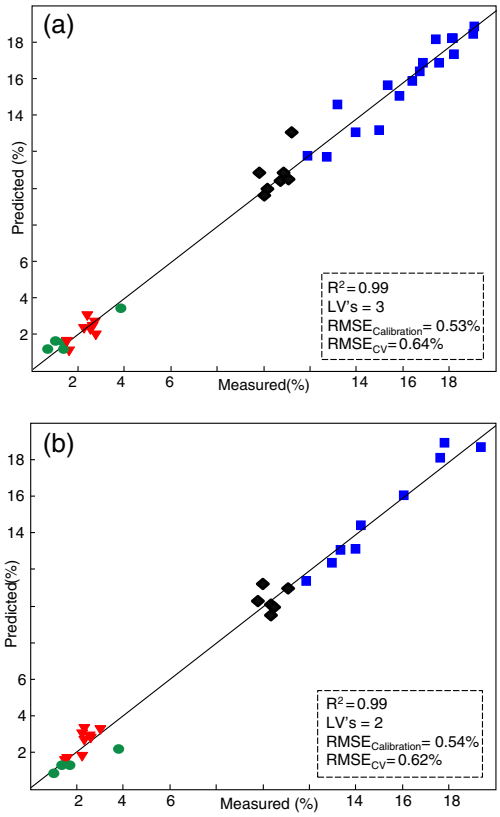


Fig. 4. Model performance of (a) in-line and (b) on-line calibration; products are labeled as described in Table 1.

model was revealed to perform slightly better with a root mean square error of prediction (RMSEP) of 0.83% as compared to an RMSEP of 0.98% for the on-line system. The increase in prediction error was not too drastic for either of the models when compared to the calibration performance (see Fig. 5 and Table 2). Only the samples positioned near the concentration gap (product type D, protease) show a bigger than expected prediction error for the on-line implementation.

Hotelling's T^2 and residuals (Q) are summary statistics which help explain how well a model is describing a given sample. Hotelling's T^2 values represent a measure of the variation in each sample within the model, whereas Q residuals are a measure of the difference between a sample and its projection into the factors (in PLS called LVs) retained in the model. In Fig. 6, test and outlying samples have been projected into the calibration model (for clarity, some of the most striking outliers have been removed). For detection of abnormal situations in the remainder of this paper, the limits corresponding to 95% coverage have been selected. This corresponds to $Q = 4.0$, $T^2 = 9.15$ for the in-line model, and $Q = 1.3$, $T^2 = 7.00$ for the on-line model.

3.3. Monitoring of the Ultrafiltration unit

Once the real-time data is available, an immediate view on the process performance and behavior is enabled. Potentially, a fully automatic and self-regulating production control system can be implemented with the monitoring equipment as input signal [10], but here we are foremost interested in the performance of the two measurement systems and the dynamics of the unit operation. Figs. 7 to 10 will present the predictions from UF of different amylase and protease types from start-up of a batch until the moment right before a cleaning-in-place (CIP) sequence, providing a comparison between on-line and in-line sampling. UF of one representative batch of protease intermediate product – a product D which is specified by a moderate enzyme protein concentration – is depicted in Fig. 7. For contrast, Fig. 8 shows the UF of two batches of an amylase type A intermediate which is a product from a low concentration range. Processing conditions for different product types can vary drastically with respect to suitable pH, ionic strength, temperature and overall capacity through the UF unit (feed flow and permeate fluxes related to concentration factors). The characteristic pattern observed during a typical UF process without disturbances looks like the profile seen in Fig. 7. The quality of the sampled NIR spectra for valid predictions of enzyme protein concentration is supported by the Hotelling's T^2 and Q residuals diagnostics trends with control limits set as determined during model development.

For a complex operation as UF in combination with the product as presented in Fig. 7, the target concentrate strength is first reached after some time. Prior to that, at the start-up of the process, various parameters such as feed flow, pH, temperature and conductivity are not at their optimal values. Hence, it is expected that spectra from the early stage of a batch are predicted differently than spectra from Normal Operating Conditions (NOC). The start-up period represents the non-Normal Operating Conditions (non-NOC) and samples used in calibration models originated solely from NOC. This is reflected on the diagnostic charts. For both probe arrangements, the first portions of the concentrate leaving the UF unit are characterized by higher Q's and T^2's, close to the control limits (Fig. 7b and c). This is explainable since in the beginning of a batch the first volume of the concentrate leaving the UF unit is expected to be more diluted due to the preceding CIP run. In particular, on-line predictions cannot be fully trusted at this stage—the concentrations were clearly underestimated. After this short run-in period, the UF system and NIR measurements quickly stabilize, an observation valid for both probe setups (Fig. 7a). The in-line probe and corresponding model seem to be more robust with respect to different conditions encountered during operation. This model works well for samples showing relatively high residuals, contrary to the on-line model. It is also interesting to note some periodic behavior in the Q-residuals in Fig. 7b (also observed for other batches). Based on the PLS calibration step, multivariate spectral information is turned into a univariate predictions. However, for monitoring purposes, it is sensible to keep track on the (spectral) variance not explained by the regression model. The pattern captured by the Q-residuals is hence related to something else. Interestingly, the fluctuations did not reveal themselves in any other (conventional) process measurement. Although we cannot completely exclude at this stage that the periodic behavior is process rather than NIR-instrument/interface based, it does hint at the potential of real-time measurement as a tool in process identification and optimization next to concentration predictions.

Fig. 8 depicts two separate runs of an amylase type intermediate and shows how alert the NIR predictions are towards the operational changes of the UF. Amylases are processed under considerably different production settings than proteases which enables for higher throughput (flows) through the unit. Any manipulation with a concentration factor is rapidly reflected in the predictions (Fig. 8a). The modus operandi of the UF unit is more complex for this product type. It is sometimes sufficient to start up the unit with a lower capacity which is seen in Fig. 8b. At those times when the capacity is increased, one can observe that the concentration decreases for a short period of time, for instance around 7:00 h and 13:00 h. Once the unit runs with full capacity and the batch is close to being fully processed, the operator will adjusts the set-point values. First, it is lowered to 1.5% (17:30 h), next it is lowered to 1% (19:30 h) and finally it increased back to 1.5%. Also, for this less concentrated intermediate, both setups function well (with the observation that the on-line calibration under-predicts at very low concentrations) and give the same picture on the process dynamics for this

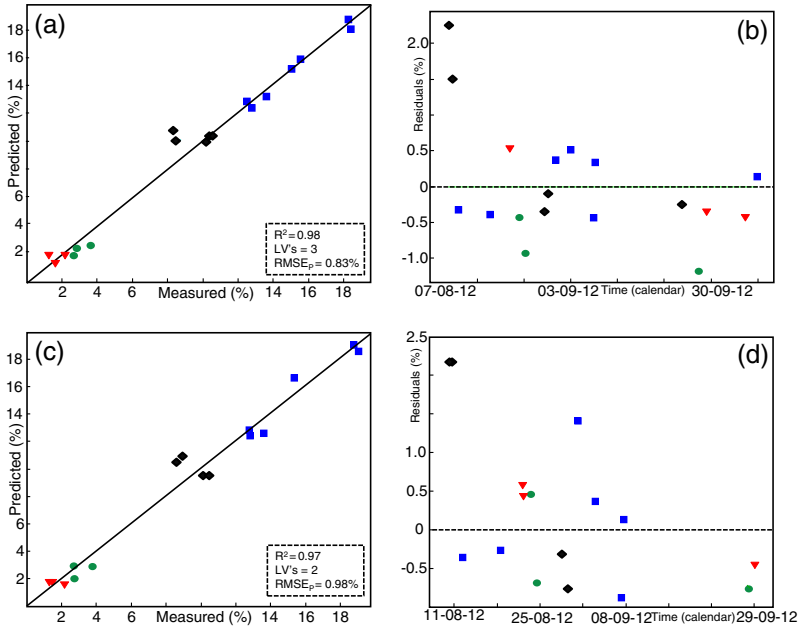


Fig. 5. Performance of the (a) in-line system with respect to reference analysis and (b) with respect to time; (c) performance of the on-line system with respect to reference and (d) with respect to time.

NOC runs. This example nicely illustrates the tremendous power of in-process measurement for visualization of what happens *inside the box* [12].

Highly concentrated protease type products sometimes lead to flux-related problems in UF operations. As this family of products can act *against themselves*, it is required to process them in the conditions unfavorable to their activity, i.e. at low temperatures and high concentrations. The challenges in purification of proteases also demonstrate itself when it comes to the arrangement of NIR probes with respect to the process stream. These issues are described in detail for two batches of type B product in Figs. 9 and 10.

Fig. 9 shows again an NOC start-up pattern which rapidly arrives at the target concentration, and the predictions are realistic and in good agreement with total dry matter measurements (not shown). The product specification requires that the UF process runs at low temperatures, and we can observe a rise in temperature of the concentrate stream happening between 14:00 h and 18:00 h for this particular run (Fig. 9b). This relatively large increase of the temperature (compared to NOC) has no significant effect on the prediction error. It was anticipated that the in-line calibration model would be robust towards the temperature fluctuations owing to the composition of the calibration set, preprocessing and spectral variable selection used [10,12]. We observed that the shape of the Hotelling's T^2 scores plotted over time follow the temperature pattern, but stayed well below the control limit (not shown). The situation in the on-line bypass-loop (Fig. 9f) is different, and, in fact, spectral residuals almost arrive at the limit at 10:30 h. Due to capacity issues upstream to the UF unit, the UF operation is put on *idle* between 10:30 h and 11:45 h. Following this break, the on-line residuals crossed the control limit and continue to increase; the on-line predictions in the last interval are not reliable (Fig. 9d). Together with the rise in spectral residuals, the predictions also slowly increased in a systematic way. This is not observed for in-line predictions (Fig. 9a)

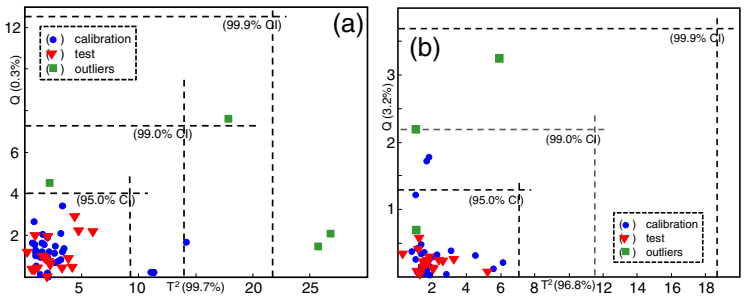


Fig. 6. Model performance of (a) in-line and (b) on-line calibration; Hotelling's T^2 vs Q residuals and their corresponding contributions including 95.0%, 99.0% and 99.9% confidence intervals (CI).

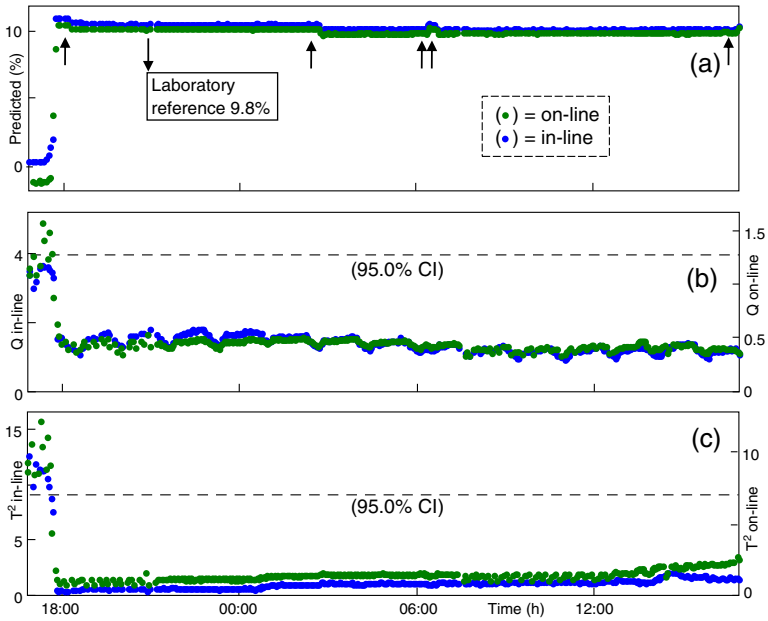


Fig. 7. Comparison between on-line and in-line NIR predictions of outcome downstream of the UF unit for a protease type product D; (a) predicted enzyme concentration, (b) spectral Q residuals; (c) Hotelling's T^2 score. Notes: arrows indicate operator set-point interventions on the concentration factor.

or dry matter measurements (not shown), suggesting an issue in the on-line sampling probe, possibly *sedimentation*.

The sampling issue is more obvious in a second example of a protease intermediate (Fig. 10, product type B). In this instance, half way into the batch, the on-line predictions run completely out of control (Fig. 10f). The situation is again preceded by a sharp increase in the spectral residuals starting around noon (Fig. 10g), accompanied by a gradual increase in (invalid) predicted concentrations (Fig. 10f). The on-line residuals peak at 15:00 h and decrease afterwards, to reach a local minimum at 17:00 h. After that, erratic fluctuations are observed, and the values for the Hotelling's T^2 statistic are following the trend (Fig. 10h). On the other hand, the in-line setup does not encounter any substantial disturbances (Fig. 10b). The in-line residuals were slightly higher than in the previously described cases in the start-up phase. This could be related to the instability in the temperature of a concentrate stream leaving the UF (Fig. 10a). When the temperature stabilizes around 11:00 h, the in-line residuals keep falling steadily for the remainder of the batch. None of the available process tags (Fig. 1) revealed patterns resembling those observed in the diagnostic plots of the on-line setup (Fig. 10g and h). However, it is clear that the flux profile (or permeate flow; Fig. 10e) was not optimal for this run from a capacity point of view. The ideal flux profile allows for a sharp decrease only at

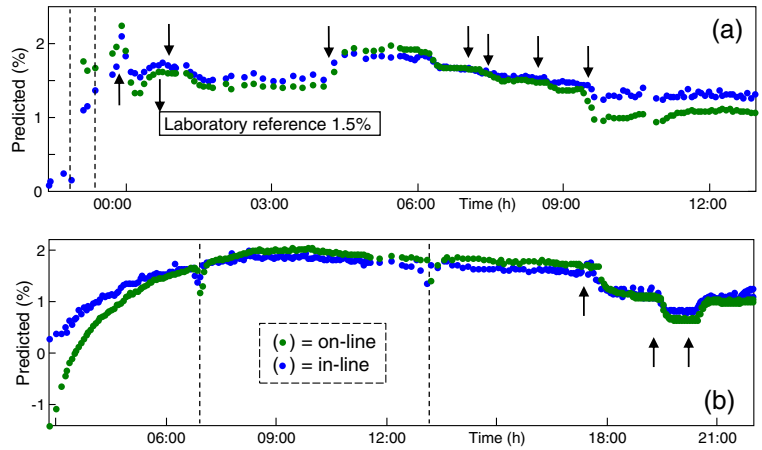


Fig. 8. Comparison between on-line and in-line NIR predictions of an amylase type enzyme intermediate product (type A). (a) batch run with full capacity from the start; (b) batch where capacity of the unit was increased gradually. Notes: vertical dashed lines mark capacity changes; arrows indicate operator set-point interventions on the concentration factor.

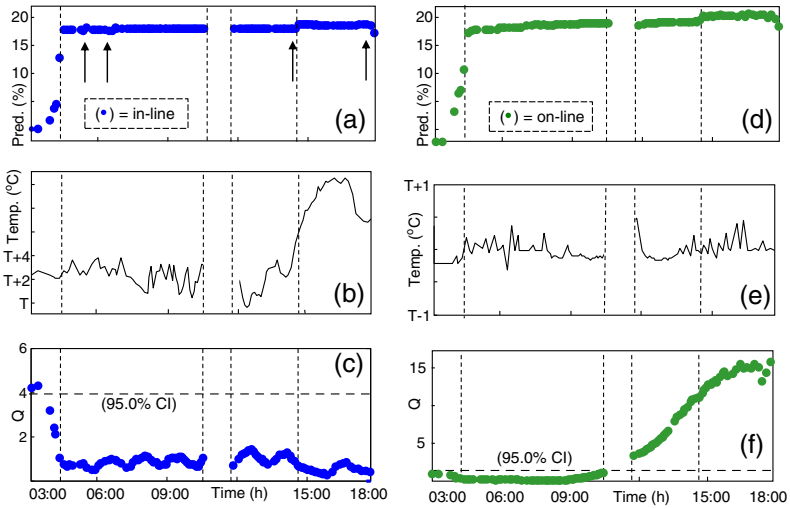


Fig. 9. (a) In-line NIR predicted enzyme concentration, (b) concentrate stream temperature, (c) in-line spectral residuals, (d) on-line NIR predicted enzyme concentration, (e) fast-loop temperature and (f) on-line spectral residuals. Notes: 10:30 h, set-point change to overcome capacity issue; 11:45 h production returns to NOC; 18:00 h production on hold due to insufficient flow.

the start up; afterwards, a steady, sufficiently high permeate flow should be reached and kept constant. This was clearly not the case for this particular batch which might indicate that membrane fouling or similar issues began early in this run. The pronounced increase in spectral residuals observed for on-line solution stems from a drastic increase in spectral baseline. An increase in scattering of the medium could indicate a phase transition in the on-line flow cell. Both the weak flux profile and the increase in on-line residuals indicate a

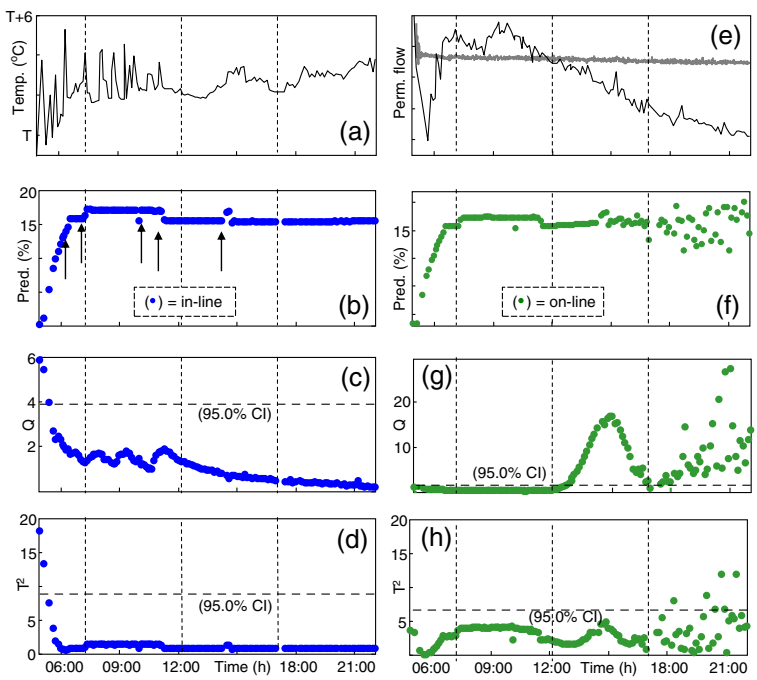


Fig. 10. (a) Concentrate stream temperature, (b) in-line NIR predicted enzyme concentration, (c) in-line spectral residuals, (d) Hotelling's T^2 scores in-line; (e) flux profiles desired (gray) and actual (black), (f) on-line NIR predicted enzyme concentration, (g) on-line spectral residuals, (h) Hotelling's T^2 score on-line.

precipitation issue on the membranes and in the fast loop, respectively. Consequently, it is speculated that the precipitate appears and leads to difficulties only in the on-line flow cell.

Protein solubility is a complex subject which involves parameters such as ionic strength, composition of the solution, pH, and temperature [13], and this was also reflected in our NIR measurement setups. At high concentrations, enzymes might undergo conformational changes, which might lead to blockage of UF membranes and serious capacity issues. This emphasizes that the production and purification of proteins is an exceptionally complex operation in which enzymes exist downstream in a *metastable* state. The observation in Fig. 10f (detected in some degree in seven out of twenty-eight production runs of product type B) speaks against the on-line setup. It is deemed that a phase transition is promoted in the on-line system mainly due to the fixed elevated temperature which is always above the processing temperature. Moreover, a *disturbance* in the form of a stopped flow can amplify the problem. To summarize, despite functioning for most batch runs, the on-line flow cell does not fully meet fundamental requirements of a proper sampling arrangement, namely not to modify the process stream or add an unacceptable delay to the response time of the system [6]. In the in-line solution, responses reflect the modes of operation of the UF unit. Without a conditioning system, there is no delay between sampling and spectra recording; hence, better traceability was achieved.

Model maintenance is very much a subject of an active investigation in process chemometrics [14–16], and operation of a process NIR-spectrometer is not maintenance-free. At the start, to improve the calibration further and gain trust of the production environment, it is recommended to continue an intensive sampling, e.g. one sample per day for an application like the one presented in this manuscript. As the implementation matures, depending on the robustness of the calibration, the frequency can be lowered. At this stage, the application can be handed over to the process operators who need to comprehend statistical process control charts for the process and NIR installation. Operators should be trained to distinguish when the alarming situation needs to be reported and if they should consult first-line support in the company or the instrument vendor. Moreover, the final users should be aware of special situations which might affect the results of chemometric models, such as: changes in the recipe, modifications to the instrument or plant setup and performance. These situations demand intensified sampling and information to local support. Periodic maintenance of chemometrics models through the data available in the process historian remains a responsibility of the local support group.

4. Conclusions

The two different NIR measurement strategies for interfacing an industrial scale UF process – on-line and in-line flow cells – both gave authentic results when compared to process and laboratory data. However, contrary to initial thoughts, the on-line flow cell was less robust during operation. The disadvantages of the latter setup include: more complex control of the sampling and conditioning system and acceleration of enzyme precipitation. The in-line system was shown to be robust towards different products (amylases and proteases) and associated processing parameters such temperature and processing speed.

Acknowledgments

The authors would like to acknowledge an Industrial PhD grant from the Danish Agency for Science, Technology and Innovation. Process engineers and technicians at the Novozymes recovery plant in Kalundborg, Denmark, are acknowledged for installing the hardware on the production line and providing invaluable advice and suggestions along the way.

References

[1] E. Baughman, Process analytical chemistry: introduction and historical perspective, in: K.A. Bakeev (Ed.), Process analytical technology, 1st edition, Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries. Wiley-Blackwell, 2008, pp. 1–11.
[2] J. Brown, A. Gilby, B. Lipták, T. Cardis, E. Baughman, Infrared and near-infrared analyzers, in: B.G. Lipták (Ed.), Instrument Engineers' Handbook: Process Measurements and Analysis, vol. 1, CRC Press, 2003, pp. 1369–1387.
[3] A.E. Cervera, N. Petersen, A.E. Lantz, A. Larsen, K.V. Gernaey, Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation, Biotechnol. Prog. 25 (6) (2009) 1561–1581.
[4] S.J. Doherty, A.J. Lange, Avoiding pitfalls with chemometrics and PAT in the pharmaceutical and biotech industries, Trends Anal. Chem. 25 (2006) 1097–1102.
[5] K.H. Esbensen, P. Paasch-Mortensen, Process sampling: theory of sampling—the missing link in process analytical technologies (PAT), in: K.A. Bakeev (Ed.), Process analytical technology, 2nd edition, Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical IndustriesWiley, 2010, pp. 37–79.
[6] M.T. Jimaré Benito, C. Bosch Ojeda, F. Sanchez Rojas, Process analytical chemistry: applications of near infrared spectrometry in environmental and food analysis: an overview, Appl. Spectrosc. Rev. 43 (5) (2008) 452–484.
[7] F. McLennan, B.R. Kowalski, Process Analytical Chemistry, Blackie Academic & Professional, London, 1995.
[8] I. Miletic, S. Quinn, M. Dudzic, V. Vaculik, M. Champagne, An industrial perspective on implementing on-line applications of multivariate statistics, J. Process Control 14 (2004) 821–836.
[9] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.
[10] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, Longman Scientific and Technical, Harlow, 1993.
[11] A. Rinnan, F.W.J. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, Trends Anal. Chem. 28 (10) (2009) 1201–1222.
[12] A.K. Smilde, F.W.J. van den Berg, H.C. Hoefsloot, How to choose the right process analyzer, Anal. Chem. 74 (13) (2002) 368A–373A.
[13] S.R. Trevino, J.M. Scholtz, C.N. Pace, Measuring and increasing protein solubility, J. Pharm. Sci. 97 (10) (2008) 4155–4166.
[14] S. Vaidyanathan, A. Arnold, L. Matheson, P. Mohan, G. Macaloney, B. McNeil, L.M. Harvey, Critical evaluation of models developed for monitoring an industrial submerged bioprocess for antibiotic production using near-infrared spectroscopy, Biotechnol. Prog. 16 (6) (2000) 1098–1105.
[15] F.W.J. van den Berg, Å. Rinnan, Calibration transfer methods, in: D.-W. Sun (Ed.), Infrared Spectroscopy for Food Quality Analysis and Control, 1st edition, Academic Press, 2009, pp. 105–118.
[16] C.B. Zachariassen, J. Larsen, F.W.J. van den Berg, S.B. Engelsen, Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amidated pectin production, Chemom. Intell. Lab. Syst. 76 (2) (2005) 149–161.

PAPER II

A. Klimkiewicz, A.E. Cervera-Padrell, F.W.J. van den Berg

Multilevel Modeling for Data Mining of Downstream Bio-Industrial Processes

Chemometrics and Intelligent Laboratory Systems (2016), *Article in press*

Multilevel Modeling for Data Mining of Downstream Bio-Industrial Processes

A. Klimkiewicz[a,b,*], A.E. Cervera-Padrell[a], F.W.J. van den Berg[b]

[a]*Novozymes A/S, Hallas Alle 1, DK-4400 Kalundborg, Denmark.*

[b]*Spectroscopy & Chemometrics section, Dept. of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.*

*Corresponding author: anna.k@food.ku.dk

Abstract

The throughput of a continuous ultrafiltration operation is limited by membrane fouling phenomena where the production capacity - monitored as flow through the membrane or *flux* - typically decreases as a function of run-time. Significant attention has been paid in both public research and industry to understanding why flux varies as these discrepancies clearly affect the production scheduling and hence economics. The potential of huge amounts of already collected full-scale processing data and its related performance history have however so far been under-employed in the production optimization efforts. The reason behind this is primarily a lack of methods capable of analyzing time series data generated during (semi-)continuous processes, characterized by widely varying operation times. The dataset examined in this investigation was compiled from records of conventional, univariate process sensors collected over several years of production of one intermediate enzyme product. Consequently, the dataset has a natural multilevel structure with level one being the process timestamps which are nested within the individual ultrafiltration runs, referred as level two. Multilevel Simultaneous Component Analysis with invariant Pattern (MSCA-P) is applied to explore this historical dataset in the context of flux decline. The unusual runs are easily identified in diagnostic plots of the between-model and the reason behind their outlying behavior becomes apparent in contribution plots. In this paper we build on the two-level idea and expand the model to a third level: *processing recipe*. MSCA-P offers a good overview during exploratory problem solving or *data mining* and helps optimization engineers to focus attention on suitable target areas. The extent of flux decline as well as higher overall mean flux value appears to be related to process temperatures and in particular increased feed temperature.

1. Introduction

Production of enzymes includes a sequence of separation, concentration, purification and stabilization steps, generally known as recovery. In modern productions, these steps generate a massive number of very diverse measurements, typically for specific and dedicated univariate monitoring and closed-loop control applications. These large amounts of data are stored in data historians but seldom used outside their direct scope, i.e. monitoring performance of individual pieces such as a particular pump or low-level control such as the temperature in a specific tank. This is a result of the demanding steps involved in thoughtful post-run data retrieval from the historian and the subsequent data pre-processing steps necessary for a more large-scale-long-term process data evaluation. There is a need for methods capable of analyzing time series data generated during (semi-)continuous processes. These types of processes show some periodicity behavior characterized by widely varying operation times, but are not easily classified as either batch or continuous in the classical sense. Consequently, the already collected full-scale process data and its related performance history are under-employed in the production optimization efforts. To change the status quo this paper discusses a chemometric method suitable for exploration of historical production records.

A great deal of correlated or redundant information is present in process measurements [1]. The information content of different process parameters also varies widely; e.g. a temperature recording in a tightly regulated tank is obviously highly relevant for the control loop but seldom useful for data mining. Therefore, even though the process database comprises of measurements on a large number of variables or *tags* (hundreds), the effective dimension of the space in which they vary in a systematic way – a state of *statistical control* with only common cause variation - is significantly smaller (usually between two and ten) [2]. Besides, often necessary information is not provided by a single process variable but rather by the way the variables are changing relative to each other or how they co-vary [1]. E.g. flows into and from the earlier mentioned regulated tank in combination with the energy demands to maintain the temperature set point might be very useful information. Latent variable (LV) methods exploit the above features of process datasets by projecting the high-dimensional data space onto the low-dimensional latent variable space. The latter represents the original data as well as possible, by accounting for the maximum amount of variance. Problems of process analysis, monitoring, and optimization are thus greatly simplified when working in this low-dimensional space of the LVs [2].

Principal Component Analysis (PCA) is a popular multivariate LV technique which has been successfully applied to analyze and monitor continuous processes [3]. In this approach, every

sampling time is represented by a row vector of a length equal to the number of process variables and the number of rows is determined by the time-horizon included in modeling. Standard PCA is however insufficient if the explored dataset includes a number of process runs (e.g. campaigns or batches). This is because it does not take into account the ordered or *blocked* nature of the data. Explicitly, standard PCA confounds the variation between and within individual data blocks. Multiway methods, on the other hand, take into account the sequential nature of data and how it is generated plus organized [1]. True batch processes can be considered as replicates of each other. Usually, a synchronization step is required after which a multiple-batch data collection forms a natural multiway structure [4]. In contrast, (semi-)continuous processes cannot easily be arranged in a similar way since they can vary considerably in duration and can be performed under quite different processing conditions. Transitions in the continuous process such as startups, grade-to-grade changeovers (set-point changes within one product type which do not require cleaning of the production line) and restarts could be considered as exceptions since they ought to follow a specific trajectory [5]. Still, runs of a *steady-state* continuous process are not suitable for synchronization since there are no *phases* or other systematic information that could be used to synchronize towards.

In this paper we want to model full process runs and we treat this type of data as a multiset/multilevel LV problem. The term *multilevel* stand for hierarchical or nested data structures and the concept of *multilevel analysis* or modeling covers a broad range of algorithms that deal with this type of data structures [6]. Data with a multilevel structure can be found in different areas of research and is most often longitudinal, multi-subject and multivariate. Methods for analysis of these types of data originate from psychometrics where scientists are interested in separating the variation inherent to the patient from the general temporal patterns [7-9]. A similar separation is often desired in metabolomics studies where systematic variability within grouped samples (e.g. metabolic biorhythms [10]) can be obscured by inter-group differences. In examples from the process industry the multilevel structure of the datasets resulted from phenomena such as catalyst deactivation (deactivation-regeneration cycles [4]) and the multi-campaign nature of a process irregularly monitored by an on-line HPLC [6]. In the above studies, variation was limited to two levels where level one constitutes measurement occasions which are nested within the individuals (subjects, patients, assessors, process runs, etc.) which can be referred to as level two. Alternatively level two is referred to as the *between* or *static* variation and level one as the *within* or *dynamic* variation. In this paper we build on the two-level idea and explore a third level, *processing recipe.* This extension has been previously suggested by de Noord [11].

The dataset examined in this investigation has a natural multilevel structure. It was compiled from several years of production of one intermediate enzyme product and consists of hundred forty-two ultrafiltration (UF) runs. As all unit operations in the recovery plant at Novozymes, Denmark, run in sequence in a continuous fashion, it is critical from a capacity-viewpoint that all of them function nearly undisturbed in accordance with the process scheduling of the full facility. However, it has been observed that the throughput of the UF varies considerably from cycle to cycle, even within the same enzyme product. The throughput is limited by membrane *fouling* or *blocking* phenomena where the production capacity - monitored as flow through the membrane, or flux - decreases over time (Figure 1). UF membranes during the steady-state filtration are mostly subjected to irreversible fouling in contrast to reversible concentration polarization building up during the startup, where the latter is easily removed during flushing [12,13]. Irreversible fouling mechanisms have been studied for the past forty years but still are not fully understood [14].

Filtration sequences or *runs* are separated by Cleaning-In-Place (CIP) operations which restore the capacity of the membranes. In daily practice, CIP is dictated either by an unacceptably low flux (one parameter in a more complex economic optimization) or because the order has been finished. As a result, high variation in filtration duration – or run-length - is experienced. This is illustrated in Figure 1 where the lengths of the shortest and longest runs in the dataset are symbolized. It can also be observed from Figure 1 that, despite differences in run-length, the characteristic developments are similar. It can be puzzling at first, as these fouling profiles resemble trajectories typically seen in batch processes. Nevertheless, the perfect development in a steady-state continuous process is expected to be a plateau, preferably situated at a high flux level. In this project it was decided to analyze the data corresponding to the (*quasi-)steady-state* UF phase. Exclusion of the startup eases the analysis since this region is abundant in nonlinearities and noisy signals. The startup finishes when the dry matter set-point is reached in the retentate stream which, from this point in time, is redirected further downstream.

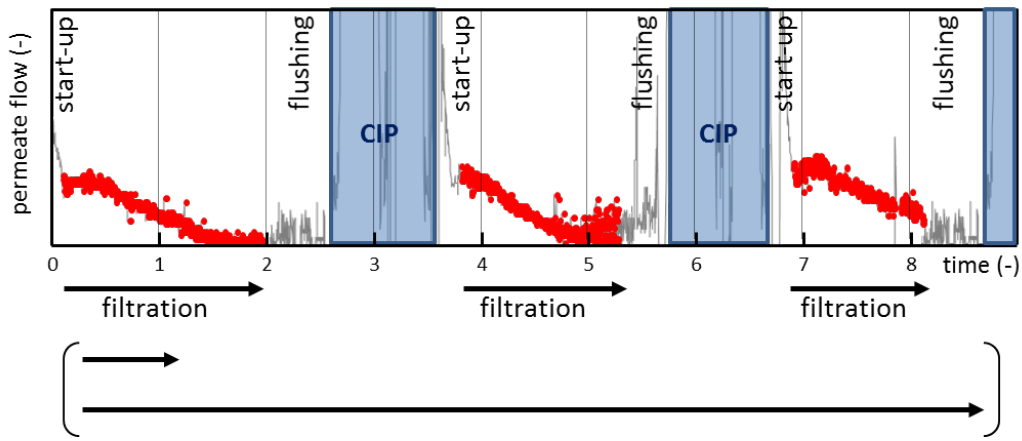


Figure 1 - Operational sequences in a continuous ultrafiltration process, shows three consecutive runs (time and permeate flow in arbitrary units).

Previous studies in flux decline were mostly univariate, done on a pilot or laboratory scale and attention was focused on the nature of the membranes, the parameters for the feed and the processing conditions. To the authors' knowledge there is up to date no multivariate study of this phenomena which utilizes only conventional, univariate process sensors. Moreover, there is a huge demand for a more efficient use of all available historical data for optimization of industrial production runs. This requires suitable data mining tools and in the case of continuous operations in downstream optimization of enzymes it is appealing to use multilevel methods. The techniques presented for data mining of process signals in the context of the flux decline problem is the subject of this study.

2. Materials and methods

2.1 Structure of the dataset

All data originates from an ultrafiltration operation in a full-scale downstream process of industrial enzymes. The process dataset is a sample from records registered over several years of production of one type of intermediate enzyme product. Each steady-state UF run i ($i = 1,\ldots, I = 142$) is represented by a data matrix $\mathbf{X}_i$ with N_i measurement occasions (timestamps) by J variables/tags ($J = 57$). The total number of timestamps over all data blocks is thus equal to $N = \sum_{i=1}^{I} N_i$. J's are average values over a fixed and equidistant time interval of conventional process measurements used for monitoring and control. They are installed at the locations depicted in the UF diagram in Figure 2. In addition to fifty-three measured process tags, four meaningful engineering parameters have been calculated based on the existing tags: volume flux, volume reduction factor, cumulated

volume reduction factor and mass throughput. Data analysis has been 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).

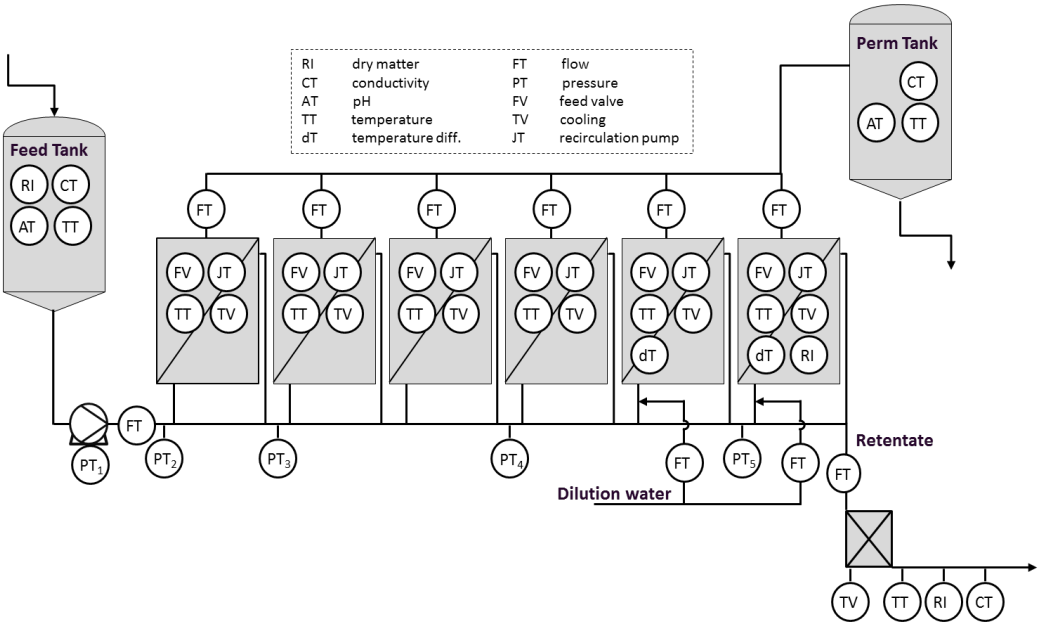


Figure 2 - Diagram of the ultrafiltration unit, built around six almost equal *UF blocks*.

2.2 Data analysis

Multilevel Simultaneous Component Analysis (MSCA) was developed by Timmerman [8] for data with a multilevel structure. It involves separate modeling of the variation between the individuals (higher level) and within the individuals (lower level). In a sense, it combines features of the factor estimation aspects in ANOVA (Analysis of Variance) and the latent variable concept of PCA [10]. The flow diagram of the least constrained MSCA-P method [7,8] – where P stands for *invariant Pattern*, the version employed in this study - is presented in Figure 3 (visualized for three runs only for simplicity).

The MSCA-P model decomposes each of the data matrices $\mathbf{X}_i$ $(N_i \times J)$ as:

$$\mathbf{X}_i = \mathbf{1}_{N_i}\mathbf{m}^{\mathrm{T}} + \mathbf{1}_{N_i}\mathbf{t}_{b,i}^{\mathrm{T}}\mathbf{P}_b^{\mathrm{T}} + \mathbf{T}_{w,i}\mathbf{P}_w^{\mathrm{T}} + \mathbf{E}_i \qquad (1)$$

where $\mathbf{1}_{N_i}$ is a $N_i \times 1$ vector of ones, $\mathbf{m}$ $(J \times 1)$ contains the offsets/average of the J process tags across all measurement occasions and all runs, $\mathbf{t}_{b,i}$ $(R_b \times 1)$ is the i-th row of the $I \times R_b$ between-runs component scores matrix $\mathbf{T}_b$, $\mathbf{P}_b$ $(J \times R_b)$ denotes the between-runs loading matrix, $\mathbf{T}_{w,i}$ $(N_i \times R_w)$ denotes the within-runs component scores matrix for run i, $\mathbf{P}_w$ $(J \times R_w)$ denotes the

within-runs loading matrix, which in the case of MSCA-P is common for all data blocks/runs, and $\mathbf{E}_i$ $(N_i \times J)$ is the matrix of residuals for run i. R_b is the number of between-runs LVs and R_w the number of within-runs LVs. Constraints are imposed so that the three parts of the MSCA-P model are orthogonal and can be solved independently [8]:

$$\sum_{i=1}^{I} N_i \mathbf{t}_{b,i} = 0_{R_b} \tag{2}$$

and

$$\left(\mathbf{1}_{N_i}\right)^{\mathrm{T}} \mathbf{T}_{w,i} = \mathbf{0}_{R_w}^{\mathrm{T}} \text{ for } i = 1, 2, \ldots, I \tag{3}$$

First, the data matrices $\mathbf{X}_i$ in our analysis are vertically concatenated in the order of their production dates (Figure 3, step 1). Because engineering variables have very diverse units and ranges the columns are scaled to unit variance jointly over all runs and timestamps (step 2). The model should focus on variation in the data and therefore the invariant part (the offset) is removed by mean centering in step 3; step 2 and 3 combined is equivalent to so-called auto-scaling the augmented data table. In step 4, average tag values for the i-th run are summarized in the i-th row vector, weighted by $\sqrt{N_i}$ and concatenated in step 5 to form the *between* data matrix ($\mathbf{X}_b$). Steps 4 and 5 are equivalent to forming the *means* data matrix ($\mathbf{X}_m$) by concatenation of i vectors of means and weighing it by the diagonal matrix $\mathbf{W}$ $(I \times I)$ with $\sqrt{N_i}$ on the diagonal. In step 6, parameters of the between-model are estimated which corresponds to estimation of a row-wise weighted PCA [8]:

$$\mathbf{X}_b = \mathbf{W}\mathbf{X}_m = \mathbf{T}_b \mathbf{P}_b^{\mathrm{T}} + \mathbf{E}_b \tag{4}$$

Weighted PCA thus takes into account the number of samples per run. Afterwards, the between-scores and residuals are back scaled according to:

$$\mathbf{T}_b = \mathbf{W}^{-1}\mathbf{T}_b \tag{5}$$

$$\mathbf{E}_b = \mathbf{W}^{-1}\mathbf{E}_b \tag{6}$$

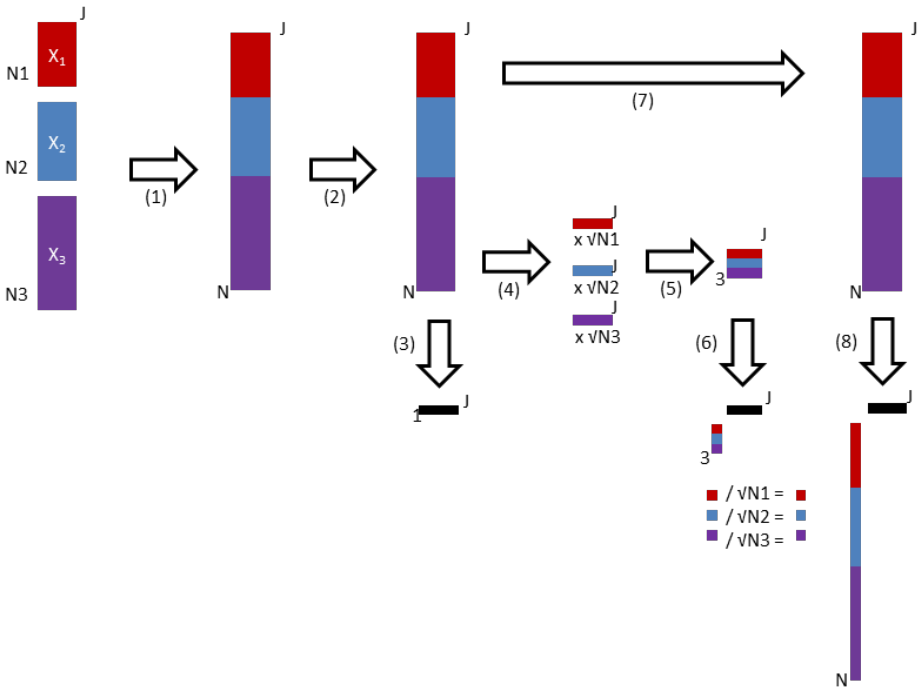


Figure 3 - Diagram of the MSCA-P method.

Solving the within part follows in steps 7 and 8 where the Simultaneous Component Analysis (SCA) idea is applied. Milsap and Meredith [15] originally developed a generalization of PCA for simultaneous analysis of a number of variables observed in several populations on several occasions. Ten Berge and coworkers [16] named the method SCA with *invariant Pattern* (SCA-P). It is equal to performing PCA (step 8) on an augmented data matrix consisting of locally (within each block/run) mean centered data blocks/runs ($\mathbf{X}_w$, step 7). Three other flavors of simultaneous component models are suggested in literature (see [7, 17]; SCA-PF2, SCA-IND, and SCA-ECP). These alternative models use constrains on the variances and co-variances of the within-component scores to impose different degree of similarity between the groups. Selection of the appropriate SCA- structure can potentially lead to a more parsimonious model which is easier to interpret [7]. Since in this project it was desirable to capture possible differences in the within-structure among the process runs we only investigated the least constrained SCA-P variant, which does not incorporate any assumptions about the relationships of the within-variation for different runs. The SCA-P models the within part of multilevel data as follows:

$$\mathbf{X}_w = \mathbf{T}_w \mathbf{P}_w^{\mathrm{T}} + \mathbf{E}_w \qquad (7)$$

Consequently, in MSCA-P it is assumed that the *within* loading matrix is invariant over all runs. The equivalence of the loading matrix $\mathbf{P}_w$ for all UF runs ensures that the components for all runs have equal interpretation and the within-scores can therefore be directly compared between different runs [8].

After removal of the offset in the two-level MSCA-P approach the preprocessed data is decomposed into two orthogonal data blocks, $\mathbf{X}_b$ and $\mathbf{X}_w$. The sums-of-squares (SSQ) of $\mathbf{X}_b$ and $\mathbf{X}_w$ can thus be used to determine the magnitudes of the within- and between-run variation in the dataset. Since each data block is reconstructed following the decomposition in Eqs. 4 and 7, respectively, it is possible to calculate the percentages of total variances taken into account by the retained R_b and R_w components/LVs by each sub-model and of the entire MSCA-P model. The variation explained by the within-components can vary considerably among the runs. It also does not automatically appear in decreasing order with respect to increasing component numbers. Calculation of the within-variation explained per each individual run can facilitate identification of the runs which do not comply well to the SCA-P solution [4,6].

In MSCA-P the same diagnostics as used in PCA are available. Hotelling's T^2 and Q-residuals are summary statistics which help explain how well a model describes a given sample. Hotelling's T^2 values represent a measure of the variation of each sample or run within the model, whereas Q-residuals are a measure of the difference between a sample and its projection onto the components retained in the model [1]. These parameters for the between-run sub-model need to be calculated taking into account the original size of the data (N) and not the number of runs (I). This is a consequence of row-weighing with $\sqrt{N_i}$ in Eq. 4. The same holds for the estimation of the other between-run model statistics such as confidence limits for scores, Q-residual and T^2. For instance, the between-level Hotelling T^2 values are calculated according to:

$$T^2_{b,i} = \mathbf{t}_{b,i}\lambda_b^{-1}\mathbf{t}_{b,i}^{\mathrm{T}} \tag{8}$$

Where the eigenvalues are determined as follows:

$$\lambda_b = eig\left(\frac{\mathbf{X}_b^{\mathrm{T}}\mathbf{X}_b}{N-1}\right) \tag{9}$$

The T^2 contributions are calculated as:

$$t_{con,i} = \mathbf{t}_{b,i}\lambda_b^{-1/2}\mathbf{P}_b^{\mathrm{T}} \tag{10}$$

3. Results and discussion

3.1 Two-level model ($R_b = 4$; $R_w = 4$)

Of the total $\mathbf{X}$ variation, after auto-scaling and before modeling, 69.0% is related to the between-run data block and 31.0% to the within-run data block; hence, both are of a considerable magnitude but between-run variance is larger. It can thus be anticipated that in a standard PCA approach on the augmented data matrix the between-variation would be dominant and it would mask the underlying patterns of within-variation. MSCA-P at two levels has been applied to describe the between-run and the within-run variation. The number of LVs in each model has been estimated by the scree procedure [18]. It appeared reasonable to use between three and four components in both sub-models. Initially, four components were selected for the between-model explaining 51.7% of the between-runs variation and the model has been explored with respect to outliers using T^2 and Q-statistics with their corresponding control limits, presented in Figure 4a. Run 12 exhibits the highest Q-residual which means that this run conforms poorly to the model. The Q-contributions of this run are plotted in Figure 4b. All variables contributing to the high residual are located on the same block of the UF unit (see Figure 2). Looking back at the raw data reveals that block E was not used during this filtration which is an unusual practice. All timestamps belonging to run 12 are also clear outliers in a four-component within-model as they exhibit significantly higher T^2 values (not shown). Consequently, run 12 was excluded, after which no strikingly outlying runs were left in the between-model with respect to Q-residuals.

A cluster of consecutive runs 48-50 separates due to high Hotelling's T^2 values (Figure 4a) as well as on the between-scores of the first up to third component (not shown). The variables contributing to the T^2 of these three runs are plotted in Figure 4c. It is clear that five process tags are jointly responsible for the high T^2 and all of them are related to the pressure control in the system. The same diagnosis could be made based on the loadings plots of components one to three where all five pressure tags are clustered, exhibiting high values that are located in the same quarter as runs 48-50. Pressure variables and the regulating pump are strongly correlated owing to a stringent control over the ultrafiltration pressure. Plotting the raw between-data (Figure 4d) confirms that these five parameters strongly follow each other as well as that they do not tend to vary except two time periods when distinct set-points were employed (runs 1-7 and runs 48-50). Pressure is the driving force in membrane filtration systems, it can however not be used to control the capacity of the UF systems because the operating pressure is factually irrelevant for the flux [19]. Since pressure in the system is maintained very accurately, any deviation around the set-point is mostly noise. Therefore, it seems that the importance of the pressure set-point changes between the runs has been

unrealistically *blown up* by variance scaling. Moreover, an examination of the loadings plot of the four-component within-model proves that there is no systematic variation to model in the pressure group (the loading values are practically zero; not shown). As a result, it was decided to lower the contribution of the pressure related variables to the overall variance in the data by removing three of the pressure tags. This was preferred over the alternative, *block-scaling* to downscale the contribution of the entire group, in order to keep model interpretation uncomplicated.

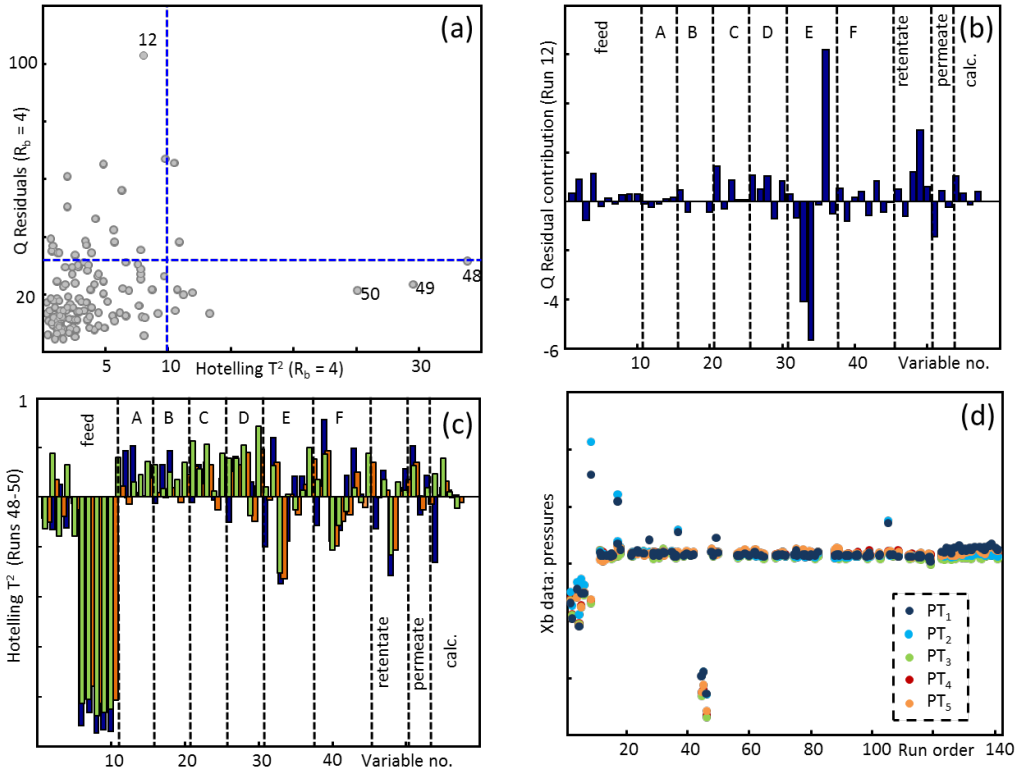


Figure 4 - Diagnostic of the outliers in the between-model. (a) Hotelling's T^2 vs Q residuals including 95.0% confidence intervals; (b) Q-residual contributions for run 12, tags are ordered according to the blocked structure introduced in Figure 2; (c) Hotelling's T^2 contribution for runs 48-50; (d) raw between-data for pressure related tags.

3.2 Two-level model (R_b = 3; R_w = 4)

The MSCA-P model has been recalculated (after removing the above mentioned runs and tags) using hundred and forty one runs, fifty-four variables and judged optimal utilizing just three between-components and four within-components. This corresponded to explaining 43.7% of the between-runs variation and 47.7% of within-variation. In total 45.0% of the variation in the auto-scaled data is explained by the two sub-models. The number of runs in the between-model with Q-residuals exceeding the 95% confidence limit is substantial (30/141; not shown). However, it is observed that addition of LVs to the between-model would serve to explain the exceptional situations rather than

common variation found on a between-run level. Namely, additional components would for instance separate clusters of runs according to the pressure step change described before. Such exceptional situations were judged irrelevant in this investigation.

3.2.1 Between-run model

The three components retained in the between-model could be interpreted based on the loadings as latent factors related to 'Process temperature' (24.7%), 'Number of blocks in use plus cooling regulation' (10.7%), and 'Mean flow rate' (7.0%).

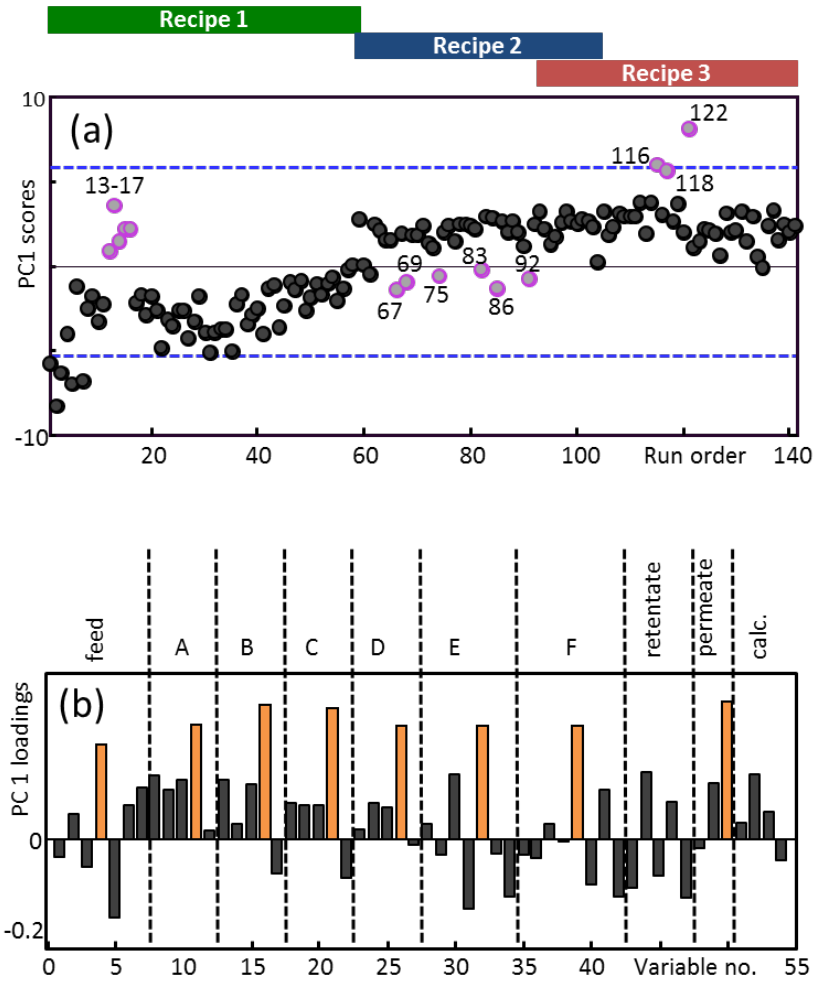


Figure 5 - First component of the between-model (R_b=3). (a) Between-run scores for component one including 95.0% confidence intervals; runs discussed in the text are highlighted and labelled; regions corresponding to the recipe changes are indicated on top of the plot; (b) Between-run loadings for component one; bars corresponding to temperature tags are highlighted.

As an example, Figure 5 presents the first PC scores ordered by production date and the corresponding loading values. Evaluation of the raw data confirms that the higher the score value the higher the processing temperature. Hence, there is a general increase with UF temperature over the investigated production years. Outliers deviating from the general trend have been highlighted in Figure 5a, and it was again confirmed that they deviate with respect to the process (set-point) temperature. Runs 13 to 17 have been manufactured using an exceptionally high temperature (a consequence of optimization trials). The most important process recipe changes are indicated on the top of Figure 5a. Recipe 2 involved a significant temperature-related change. As can be seen, most of the runs have a positive score on component one from this point in time and onwards, but there are 6 exceptions (67, 69, 75, 83, 86, 92) with negative score on component one. This appears to be because of some manual adjustments since in those runs the temperatures on the UF blocks have been put to the recipe 1 set-point. Finally, runs 116, 118, and 122 are indeed processed under particularly high temperatures.

The between-run model describes the static changes/modifications and how the ultrafiltration process is executed over calendar time. Those changes were identified predominantly as deliberate engineering input to the process recipe plus some instances of operators' interventions driven by the optimization efforts.

3.2.2 Within-run model

Four within-run components have been retained in the model explaining 24.7%, 10.7%, 7.0% and 5.3% of the within-run variation, which sums up to 47.7%. Inferring from the loadings these four components are interpreted as latent factors related to 'Permeate flux/Throughput', 'Feed flow and block addition/removal', 'Conductivity' and 'Cooling regulation'. Therefore, the highest amount of variation occurring systematically over the course of filtration is related to flow through the system.

Static parameters governed directly by the set-point values fixed once per run are now positioned close to zero in the loading plots. Signals related to pressure, temperature, pH or dry matter readings belong to this group. In contrast to the between-model the within-run model focuses on process parameters which change dynamically over the course of an ultrafiltration run. It is therefore convenient to investigate how the latent factors evolve - that is, to plot the within-run scores as time series.

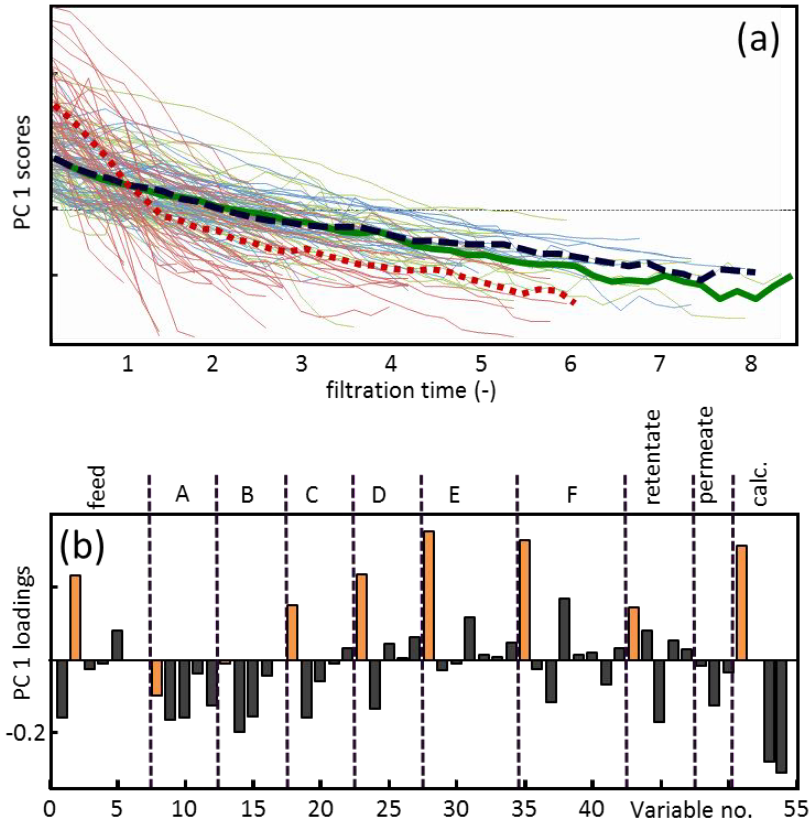


Figure 6 – First component of the within-model (R_w = 4). (a) Within-run scores for component one plotted as time series; the thin lines in the background represent the true profiles of individual runs colored according to the recipe variant; thick lines represent the mean for all available data for each given time-point for each recipe (green solid line- recipe 1; blue broken line – recipe 2; red dotted line – recipe 3) (b) Within-run loadings for component one; bars corresponding to flow tags are highlighted.

Figure 6a shows these time series plotted per run for the first component. Lines are color coded according to the recipe version (as was introduced in Figure 5). The mean flux trajectory per recipe is depicted by bold lines. Unsurprisingly, the throughput decreases over filtration duration for all runs. However, the trajectory for the newest recipe is clearly different from those experienced in the older versions. The profiles covered by the first component are in a good agreement with what is seen in the readings from a number of flowmeters in the real process. Therefore, the distinct shape of the mean flow trajectory of recipe 3 could be interpreted as follows: flux/throughput is significantly higher at the start, but declines faster than what is observed for the earlier recipes. Additionally, the new-recipe runs are significantly shorter. This brings the conclusion that the membranes get fouled/blocked faster when the UF runs according to the new recipe and as a consequence this calls

for a more frequent (sometimes unforeseen) cleaning. This cleaning requires high volumes of water and chemicals but primarily a considerable amount of time [20].

3.3 Three-level model ($R_r = 2$; $R_b = 2$; $R_w = 4$)

Separation has been observed on both levels of the MSCA-P model which has been attributed to the process recipe valid at that point in time. This was a motivation to expand the MSCA-P model into a three-level structure using the optimized dataset with fifty-four variables as in the preceding section. Explicitly, the variation in the auto-scaled data has been split into three levels which are modeled separately. This also includes an additional stage in the procedure depicted in Figure 3 which enters between step 3 and 4. First, the average tag values of each of the three recipes are calculated and weighted by $\sqrt{N_r}$ where N_r represents the number of time points when a specific recipe has been in force. Next, the *processing recipe* data matrix ($\mathbf{X}_r$) is formulated by concatenation of – for our case - three vectors of weighted means. Parameters of the recipe-model are estimated by PCA using R_r components. Scores and residuals per recipe are back scaled with the inverse of $\sqrt{N_r}$. Next, the overall data is locally mean-centered by the recipe means and the procedure continues, as previously described, with steps 4-8 (Figure 3). The three constructs of the auto-scaled data blocks ($\mathbf{X}_r$, $\mathbf{X}_b$ and $\mathbf{X}_w$) are orthogonal to each other as before and their SSQs add again up to 100%.

The overall statistics of the data to which the three-level MSCA-P has been applied is shown in Figure 7. After the invariant part is removed, 13.3% of the variation is related to the between-recipe level, 53.8% to the between-run level, and 32.9% to the within-run variation. Part of the variance previously found at the between-level is now moved to the recipe-level, which also affects the number of components used to explain the systematic part of variation present in the between-run part. Specifically, information related to the 'Mean flow rates' and part of the 'Temperature related information' is now covered by the between-recipe level. The outcome of the between-level is slightly different but has exactly the same interpretation with component one related to 'Number of blocks in use plus cooling regulation' and component two to 'Process temperature' (not shown). On the other hand, the within-variation is completely unaffected by the addition of the extra level to the model, which is expected as the original two sub-models are orthogonal and hence do not affect each other. Also the total variation explained by the three sub-models is the same as in the two-level model and adds up to 45.0%. Figure 7b also emphasizes that the within-model variance captured by the MSCA-P model, characterized by one common loading base $\mathbf{P}_w$ in Eq. 7, can differ significantly between runs [8].

3.3.1 Between-recipe level

Figure 8 presents the parameters of the between-recipe model. The two-component model fully explains the data, and this plot indicates what makes the three recipes different. In agreement with previous observations, recipe 1 differs in processing temperature from recipe 2 and 3, which is explained by the first component. The second PC separates recipe 2 from recipe 3 primarily owing to the temperature of the feed, which has the highest loading on this PC. This is the main candidate parameter that needs to be investigated in context of a faster flux decline in the recent UF runs. However, at the same time it appears that the new recipe runs exhibit the highest mean flux value which is a desirable feature. Other differences between recipes can also be derived from Figure 8. For instance, recipe 1 differs from succeeding recipes with respect to dry matter content in the feed and recipe 2 is characterized by the distinct dry matter set point in the retentate.

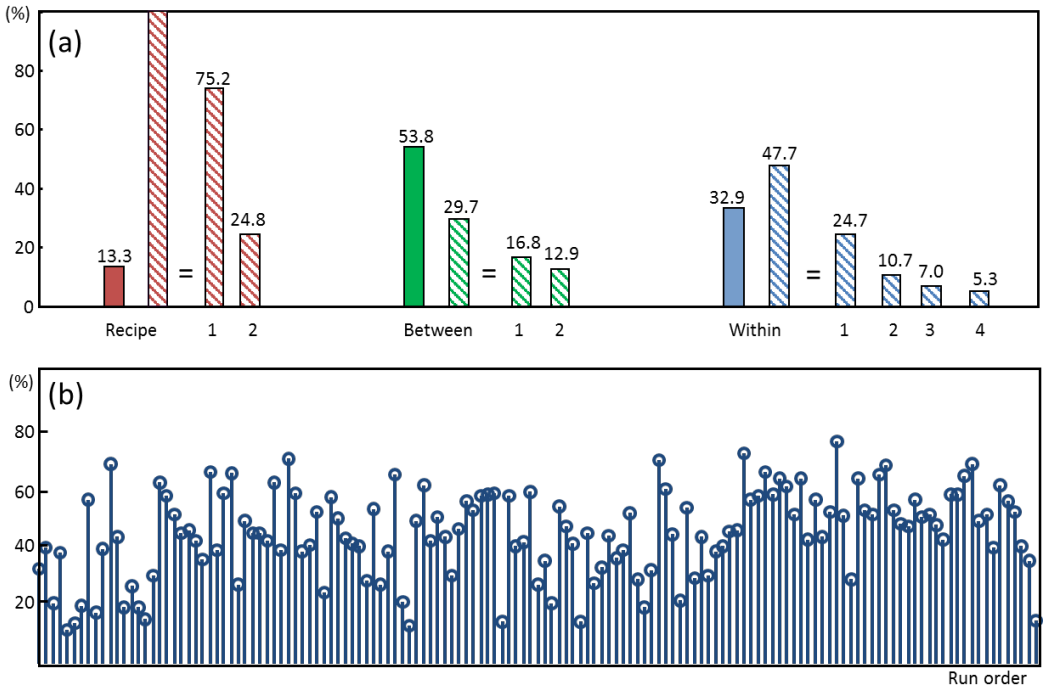


Figure 7 – Statistics related to the auto-scaled data decomposed, modeled at three levels (R_r = 2; R_b = 2; R_w = 4); (a) magnitude of the variation contributed to a specific level is represented by the solid bars, percentage of variance explained by the corresponding sub-models and per component retained in the model is indicated with patterned bars; (b) total (R_w = 4) within-run variance explained in each run.

Concentration and purification of enzymes is a complex process where numerous factors have to be taken into account to secure a profitable outcome. The process needs to be optimized with respect to enzyme solubility and stability. Proteins are known to have an extreme effect on flux and separation properties of UF membranes. This effect is related to both the micro-environment (ionic

strength, pH, etc.) and the processing conditions [21,22]. Density and viscosity of the liquid increases with the protein concentration [21], and with highly viscous liquids it is harder to obtain sufficient cross-flow over the membrane area and good fluxes. This issue is often addressed by increasing the feed temperature which decreases the viscosity of the liquid and thus leads to a better mass transfer in the feed and consequently to a higher flux [19]. Indeed, the flux at the start of the process is higher in runs produced according to recipe 3 (Figure 6a) characterized by distinct (higher) temperature of the feed (Figure 8). On the other hand, at high concentrations proteins may undergo conformational changes owing to either increased exposure of the membrane to the protein or due to the high shear rates on the membrane surfaces [22]. This would lead to increased aggregation, protein deposition and clogging of membranes. It is suspected that these adverse phenomena take place at a higher rate when the high throughput is obtained by increasing the temperature of the feed and on the UF blocks. In addition, pH of the solution affects interactions between proteins and between proteins and the membrane. This parameter also appears to make the distinction between recipes 2 and 3 (Figure 8, PC2). Ultimately, designed experiments that incorporate the above findings should be conducted in full-scale production (or alternatively a recovery pilot plant) to find the right direction for further process optimization.

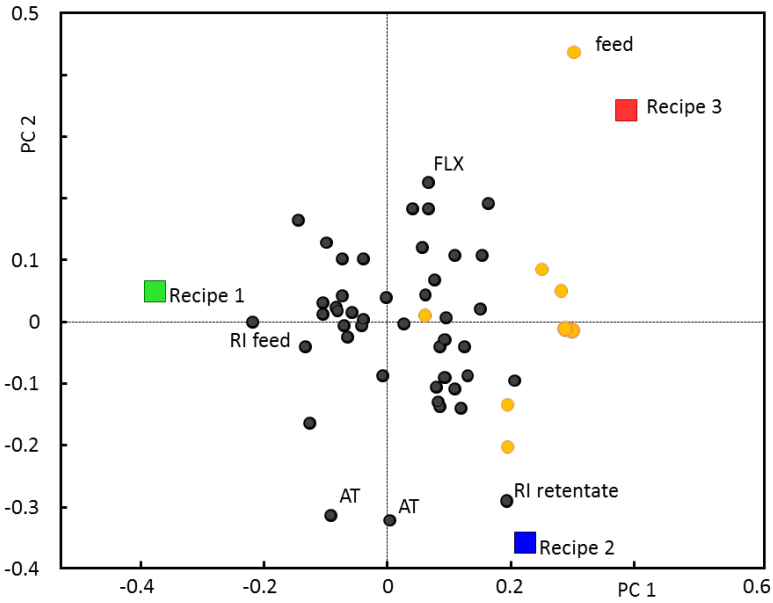


Figure 8 – Bi-plot of the between-recipe model (R_r = 2), scores (squares) and loadings (dots); loadings of temperature tags are highlighted and loadings of the tags with the highest influence on the model labelled.

3.3.2 Within-run level

The within-model variation is exactly the same as for the two-level approach previously described. It is worth noticing that the variation explained by the four within-components varies considerably for individual runs (Figure 7b). In terms of within-component 1, previously connected with 'Permeate flux/Throughput', the explained variance per run (Figure 9a) is significantly higher in runs produced according to recipe 3. As reported in literature [4,6], in case of a continuous processes small amounts of variance explained per run does not automatically lead to a negative interpretation, since ideally continuous production should be a steady-state process. In this study, forty-one runs have a low (<10%) within-variance explained by the first component (permeate-flux related). All of them exhibit reasonably stable, slowly decreasing profiles (see Figure 9b, run 5 as an example). On the other hand, runs having the highest variance explained typically show a steep, fast declining profiles on within-PC1 (see for instance run 120, Figure 9b).

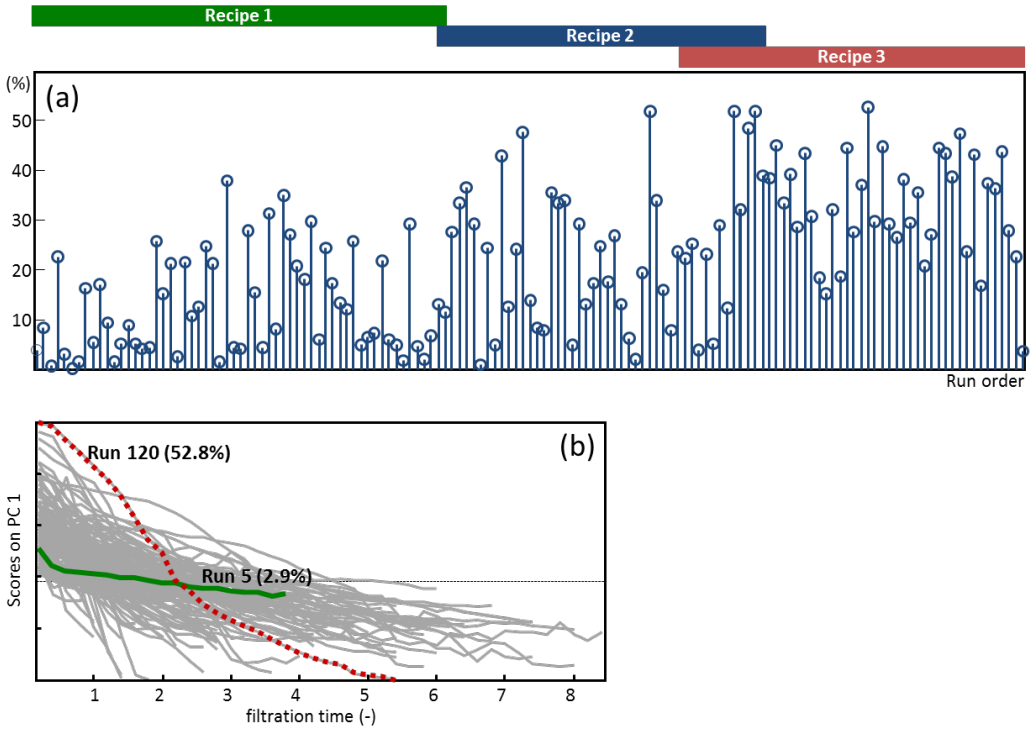


Figure 9 - Within-run variance explained by the first component for each run (a); selected profiles, run 5 and run 120.

4. Conclusions

MSCA-P has proven to be a powerful method when studying data of a (semi-)continuous process collected over a large time span, run under different recipes and including experimental runs (in the

form of optimization trials and operator interventions). The latent behavior of the ultrafiltration system investigated in our work has been modeled at three levels. The underlying factors behind recipe-, between- and within-models have been identified which greatly eased the exploration of the fifty-seven parameters used to investigate the process. The unusual runs are easily identified in diagnostic plots and the reason behind their outlying behavior explained in contribution plots. Parameters related to the flux decline have been found on the recipe-level and the within-runs level. MSCA-P applied on these large amounts of production data offered a good overview during exploratory problem solving and helps optimization engineers to focus attention on suitable target areas. The steep flux decline as well as higher overall mean flux value appears to be related to process temperatures and in particular increased feed temperature. This study revealed that higher processing temperatures can lead to both positive and negative effects in the examined membrane separation system.

Future work of the optimization engineers should be focused on finding the balance between the UF temperature and the parameters of individual UF orders (feed volume, degree of concentration, etc.). Especially, the question should be addressed if, from an economic perspective, it is better to process the same order in several high throughput runs (involving extra cleaning) or in one extended run (without the need for extra CIP).

Acknowledgements

The authors would like to acknowledge an Industrial PhD grant from Innovation Fund Denmark. Process engineers, scientists and technicians at the Novozymes recovery plant and optimization group in Kalundborg, Denmark, are acknowledged for providing invaluable advice and suggestions along the way.

References

[1] B.M. Wise, N.B. Gallagher, The process chemometrics approach to process monitoring and fault detection, J. Process Control 6 (1996) 329-348.

[2] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, Int J Adapt Control Signal Process 19 (2005) 213-246.

[3] J.V. Kresta, J.F. Macgregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, The Canadian Journal of Chemical Engineering 69 (1991) 35-47.

[4] O.E. de Noord, E.H. Theobald, Multilevel component analysis and multilevel PLS of chemical process data, J. Chemometrics 19 (2005) 301-307.

[5] T. Kourti, Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions, J. Chemometrics 17 (2003) 93-109.

[6] D.L. Ferreira, S. Kittiwachana, L.A. Fido, D.R. Thompson, R.E. Escott, R.G. Brereton, Multilevel simultaneous component analysis for fault detection in multicampaign process monitoring: application to on-line high performance liquid chromatography of a continuous process, Analyst 134 (2009) 1571-1585.

[7] M.E. Timmerman, H.A. Kiers, Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences, Psychometrika 68 (2003) 105-121.

[8] M.E. Timmerman, Multilevel component analysis, Br. J. Math. Stat. Psychol. 59 (2006) 301-320.

[9] E. Ceulemans, M. Hubert, P. Rousseeuw, Robust multilevel simultaneous component analysis, Chemometrics Intellig. Lab. Syst. 129 (2013) 33-39.

[10] J.J. Jansen, H.C. Hoefsloot, J. van der Greef, M.E. Timmerman, A.K. Smilde, Multilevel component analysis of time-resolved metabolic fingerprinting data, Anal. Chim. Acta 530 (2005) 173-183.

[11] O.E. de Noord, Multivariate analysis and monitoring of dynamic process data, 25th Annual Symposium on Chemometrics, Dutch Chemometrics Society, Delf, the Netherlands (2009).

[12] G. van den Berg, C. Smolders, Flux decline in ultrafiltration processes, Desalination 77 (1990) 101-133.

[13] V. Gekas, Terminology for pressure-driven membrane operations, Desalination 68 (1988) 77-92.

[14] J. Linkhorst, W.J. Lewis, Workshop on membrane fouling and monitoring: a summary, Desalination and Water Treatment 51 (2013) 6401-6406.

[15] R.E. Millsap, W. Meredith, Component analysis in cross-sectional and longitudinal data, Psychometrika 53 (1988) 123-134.

[16] J.M. Ten Berge, H.A. Kiers, V. Van der Stel, Simultaneous components analysis, Statistica Applicata 4 (1992) 377-392.

[17] M.E. Timmerman , E. Ceulemans , A. Lichtwarck-Aschoff, K. Vansteelandt, Multilevel simultaneous component analysis for studying intra-individual variability and inter-individual differences. In: Dynamic process methodology in the social and developmental sciences, 2009 (pp. 291-318), Springer US.

[18] R.B. Cattell, The scree test for the number of factors, Multivariate Behavioral Research 1 (1966) 245-276.

[19] J. Wagner, Membrane Filtration Handbook: Practical Tips and Hints, Osmonics Minnetonka, MN, 2001.

[20] J.K. Jensen, J.M. Rubio, S.B. Engelsen, F. van den Berg, Protein residual fouling identification on UF membranes using ATR-FT-IR and multivariate curve resolution, Chemometrics Intellig. Lab. Syst. 144 (2015) 39-47.

[21] M. Cheryan, Ultrafiltration and Microfiltration Handbook, Technomic Pub. Co, Inc., Lancaster (1998) 264.

[22] A. Marshall, P. Munro, G. Trägårdh, The effect of protein fouling in microfiltration and ultrafiltration on permeate flux, protein retention and selectivity: a literature review, Desalination 91 (1993) 65-108.

PAPER III

A. Klimkiewicz, A.E. Cervera-Padrell, F.W.J. van den Berg

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

Industrial & Engineering Chemistry Research (2016), *Submitted*

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

Anna Klimkiewicz [a,b,*], Albert E. Cervera-Padrell[a], Frans W.J. van den Berg[b]

[a]Novozymes A/S, Kalundborg, Denmark;

[b]Spectroscopy & Chemometrics section, Dept. of Food Science, Faculty of Science, University of Copenhagen, Frederiksberg, Denmark.

*Corresponding author: anna.k@food.ku.dk

Abstract

This study investigates flux decline in ultrafiltration as a capacity measure of the process. A continuous ultrafiltration is a multi-stage 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) 'physical location' of the sensors on the process layout, and 2) 'engineering type of tags'. Abnormal runs are removed iteratively from the original dataset, 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 has an optimization potential.

Keywords: multivariate modeling; multiblock PLS; latent variables; ultrafiltration flux; membrane fouling

1. Introduction

Classical engineering strategies do not always perform well in control and optimization of full-scale bio-manufacturing 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 concepts1. 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 like PCA and PLS. The popularity of multiblock methods has however grown only modestly over time. An explenation 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 datasets [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 dataset [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] sector. 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 bio-processing [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 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].

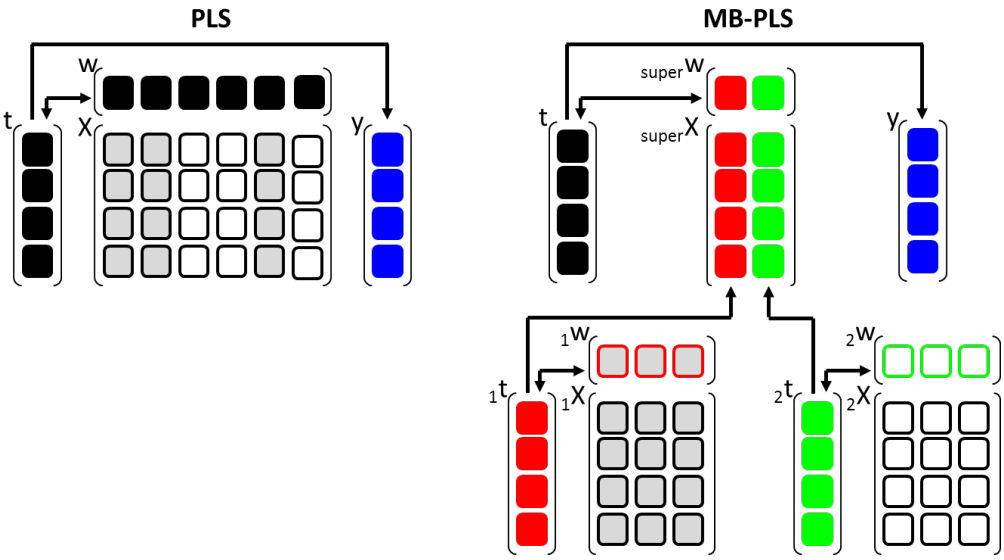


Figure 1. Conceptual scheme of the MB-PLS model.

There are three prevalent ways to obtain MB-PLS models [10]. The first method uses the block-level scores for deflation of $\mathbf{X}$ and $\mathbf{y}$ [11] which ensures orthogonality between the block-level scores. In the second approach, the algorithm uses the super-level scores to deflate $\mathbf{X}$ and $\mathbf{y}$ [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 one 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 the missing values. In the third method, only $\mathbf{y}$ 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 or 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 incorrect [5,6].

This study uses flux in ultrafiltration (UF) as a capacity measure of a process and focusses on the block level to investigate the overall flux values. 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 bio-manufacturing. 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 multi-stage 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 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.

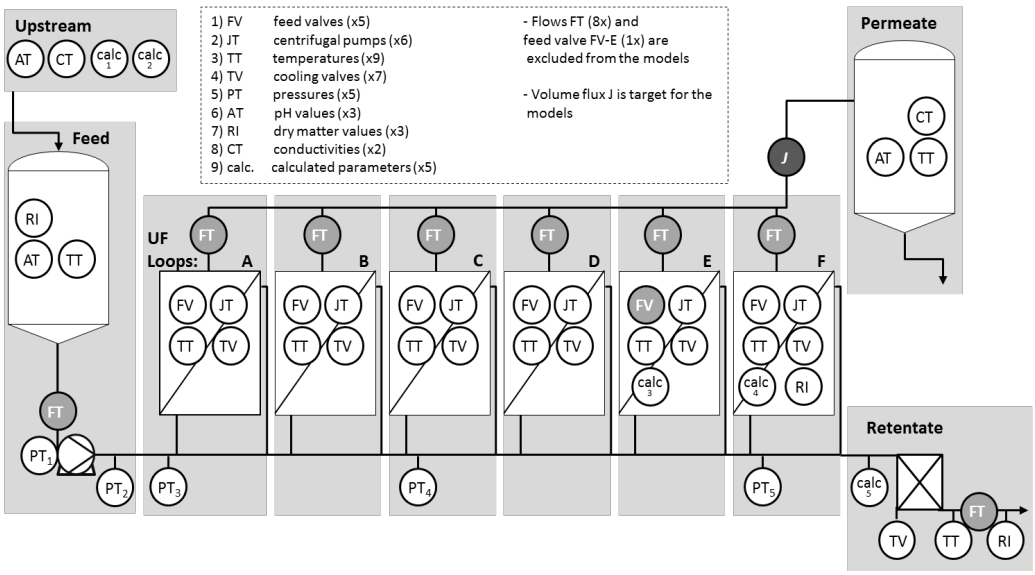


Figure 2. Schematic of the ultrafiltration system plus the approximate location of forty-nine process measurements and five calculated engineering values; flow signals and the throttling valve on loop E are excluded during modeling.

A membrane system designed as multi-stage 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 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, process temperatures, properties of the feed, 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 unacceptable low values. It is, however, problematic to use these seven parameters for the post-run 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:

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

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 number of loops working at timestamp t, based on assumption that a loop is working if the power of the corresponding centrifugal

pump is larger than 1%, and A is 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 dataset

All data originates from an ultrafiltration operation in a full-scale downstream process of industrial enzymes. The process dataset 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 [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 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 dataset. 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 ($i = 1,\dots, I$) 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 datasets is equal to $N = \sum_{i=1}^{I} N_i = 623$. Process measurements recorded upstream have been stretched or extrapolated 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 in the examined dataset. 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].

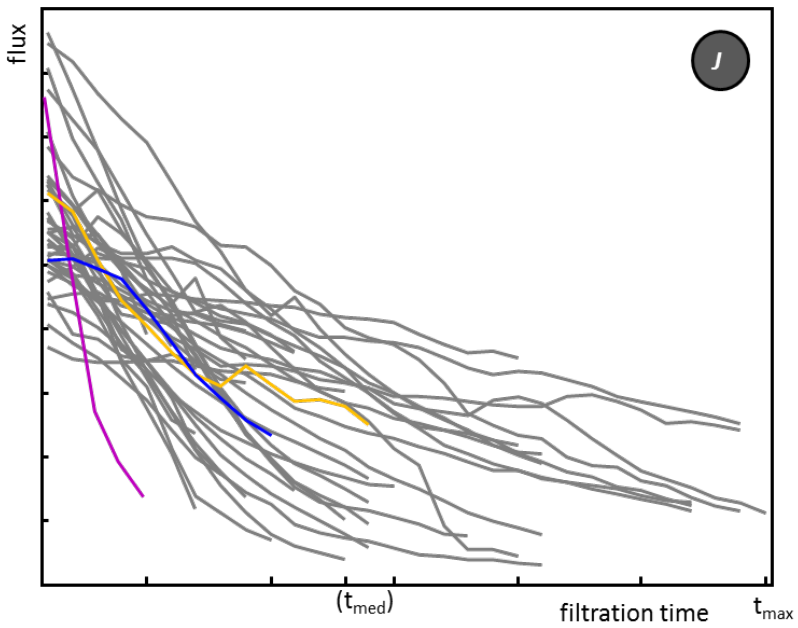


Figure 3. Flux values/fouling profiles encountered in the investigated dataset (I = 40).

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, 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 Figure 2, comprising between two and eight 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]. 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 forty runs has been concatenated in the process tag direction (variable-wise), and all data has been auto-scaled. A PLS model is built between the process variables $\mathbf{X}$ $(N \times J)$ and the flux $\mathbf{y}$ $(N \times 1)$. 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 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.

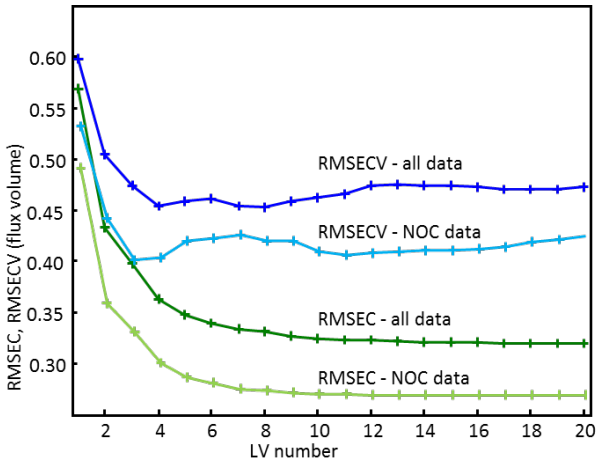


Figure 4. Calibration (RMSEC) and cross-validation (RMSECV) errors for the PLS models constructed using all data (I = 40, N = 623) or NOC data (I = 30, N = 508).

Stratified cross-validation has been applied to determine the optimal number of latent factors in the model where each of the investigated runs 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 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].

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 $\mathbf{y}$ (68.8%) and the subsequent components explain significantly less variation of the flux (13.1%, 3.4% and 1.9%, respectively).

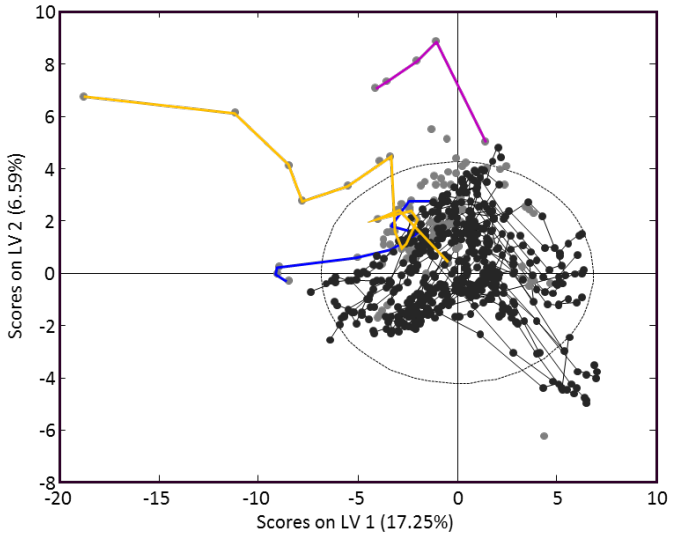


Figure 5. LV1 vs. LV2 score plot of the PLS model constructed on NOC data; three instances of AOC runs are projected onto the model, blue and yellow represent abnormal pressure behavior, purple shows abnormal temperature behavior (see text for details).

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 behavior of 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 dataset, 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 only NOC runs. A new PLS model was calculated using the remaining thirty runs. The stratified cross-validation procedure points at three LV's 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 $\mathbf{y}$ and 35.9% of the variation in $\mathbf{X}$.

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 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 case of the abovementioned 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 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.

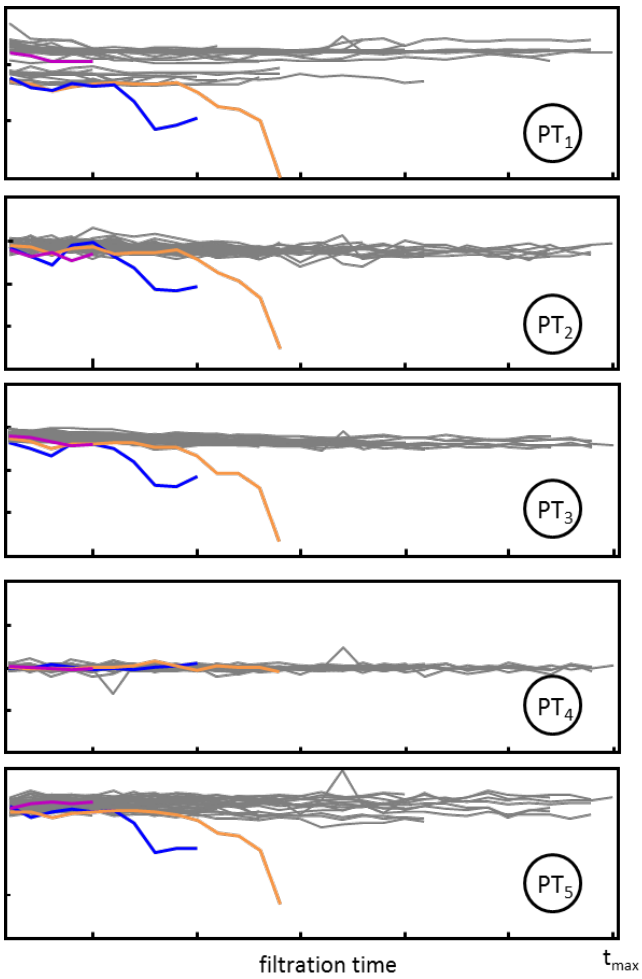


Figure 6. Signals related to pressure control (see Figure 2) in the UF system collected during NOC runs (marked in gray) and examples of the AOC runs, pressure related (blue and yellow) and temperature related (purple; compare with Figure 5).

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 (thirty runs, three LV's). 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). 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 auto-scaling preprocessing).

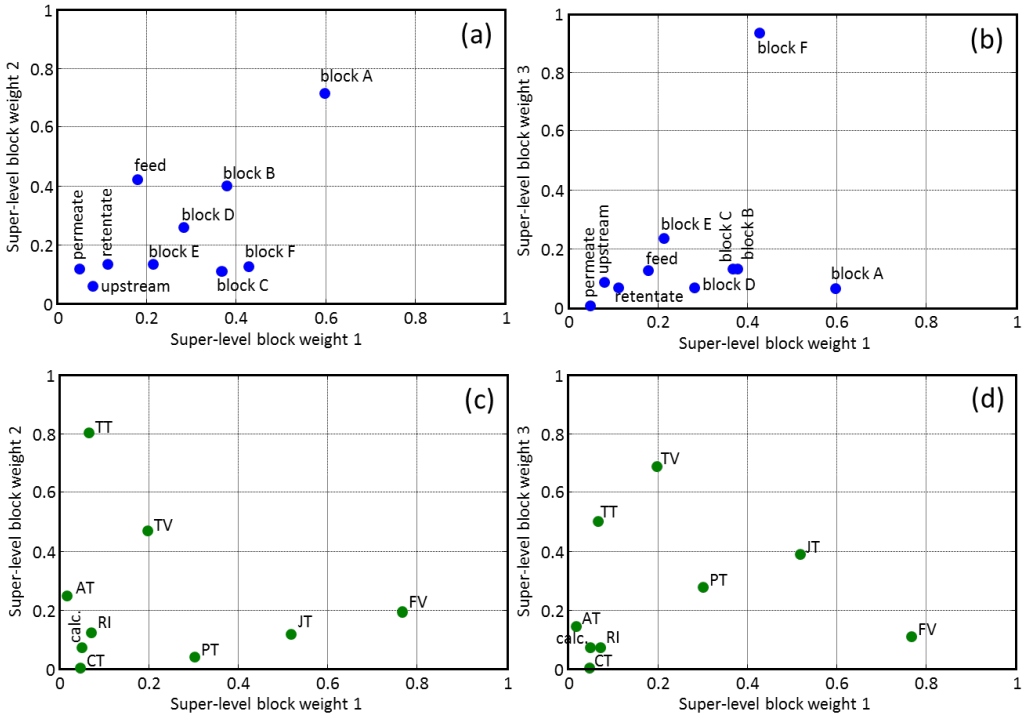


Figure 7. MB-PLS model super-level block weights when blocking is done according to (a-b) physical location or (c-d) sensor type.

Figure 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 two blocking strategies.

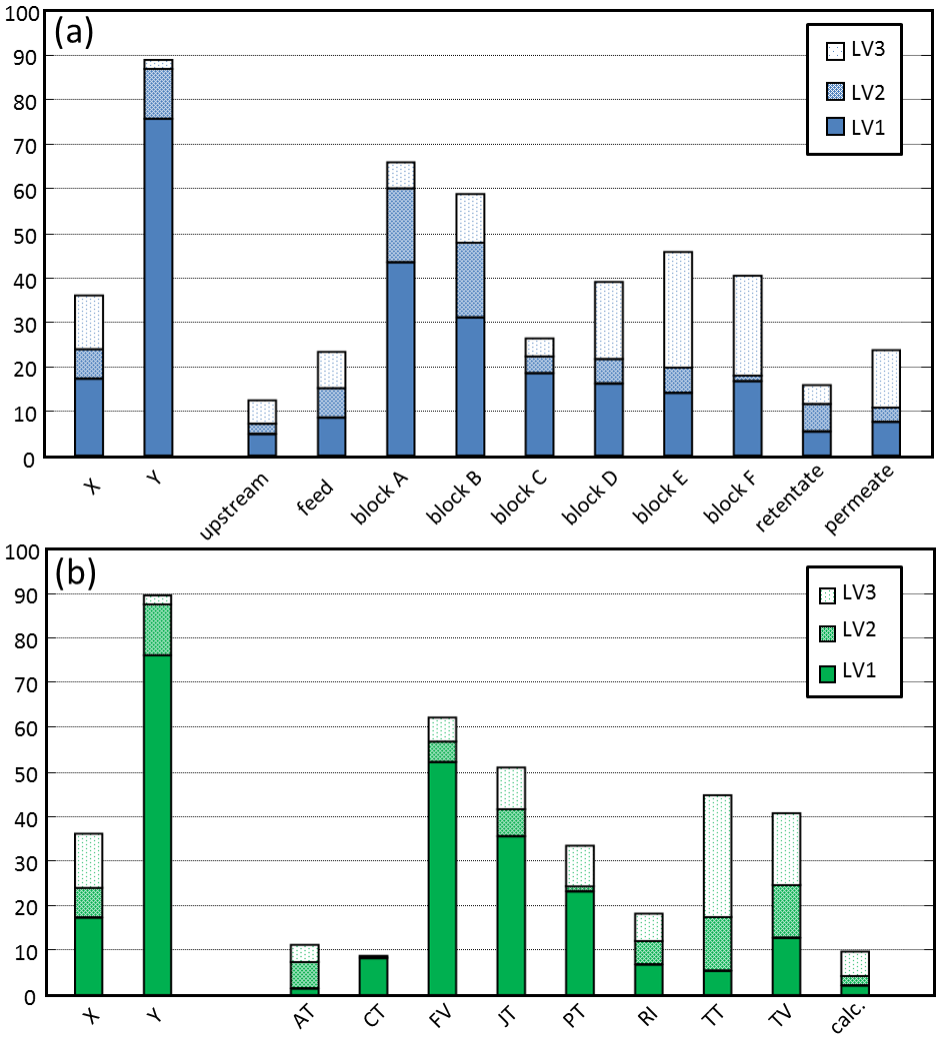


Figure 8. MB-PLS model variances explained when blocking is considered according to (a) physical location or (b) sensor type (see Figure 2 for interpretation).

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 predicted or 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 knowledge, 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 LV's 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 LV's 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 dataset where pressure tags quickly fall out of analysis as they appeared constant over the course of ultrafiltration [14].

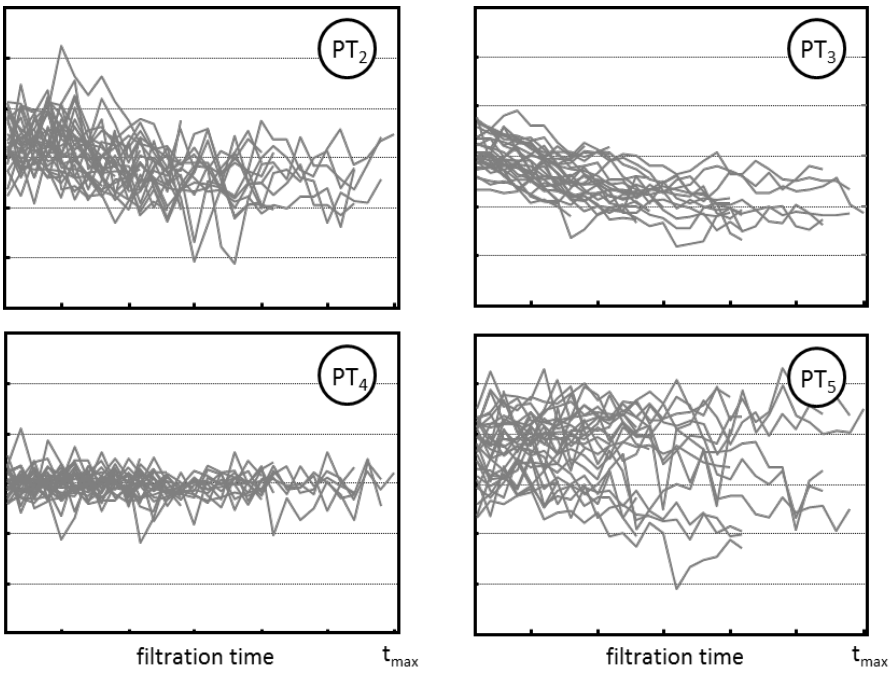


Figure 9. Pressure measured during the NOC runs (see Figure 2).

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 on 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 LV's (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 models and the low-level interpretation of the raw data. If we compare the mean flux trajectories up to the median filtration length of the investigated dataset, it can be seen that the higher processing temperature converts into higher flux at the start of a run but a lower flux towards 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.

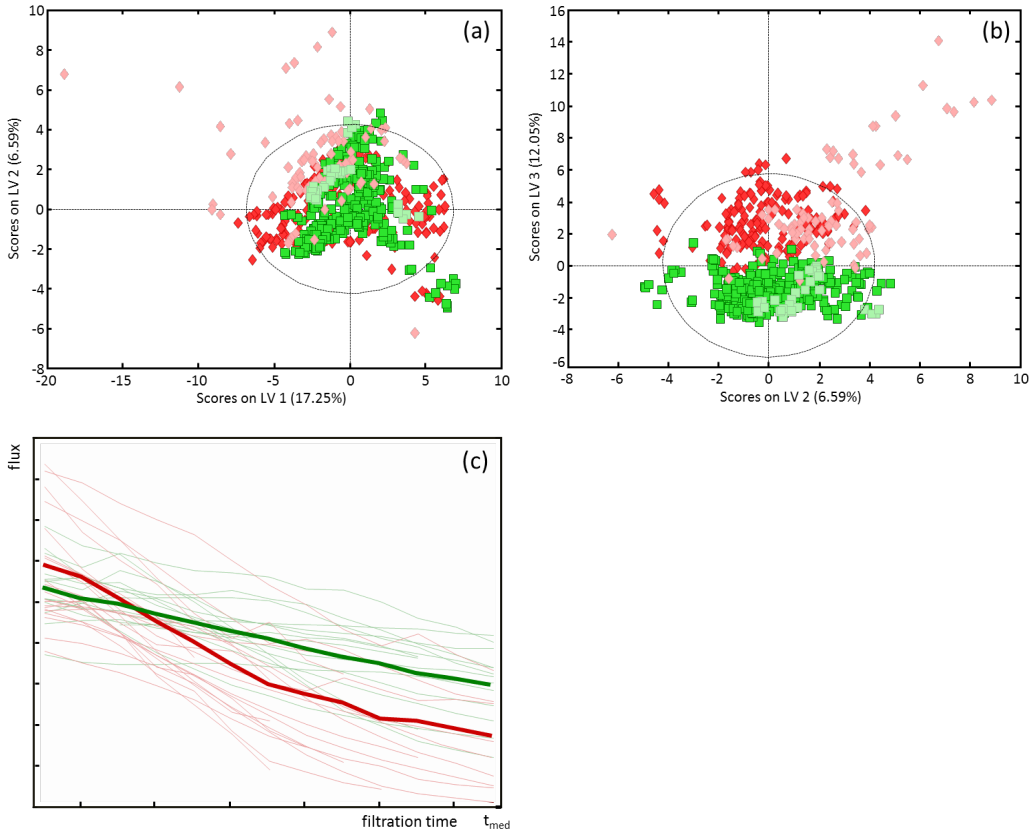


Figure 10. (a) LV1 vs. LV2 (analog to Figure 5) and (b) LV1 vs. LV3 score plots of the PLS model constructed on NOC data colored according to a temperature related processing recipe change: red markers - equal temperature on all loops; green markers - lower temperature on D-F loops; shaded markers: AOC (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 recipe.

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 type 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 analysis of the multivariate historical datasets even more.

Acknowledgements

The authors would like to acknowledge an Industrial PhD grant from Innovation Fund Denmark. Process engineers, scientists and technicians at the Novozymes recovery plant and downstream optimization department in Kalundborg, Denmark, 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. Chemometrics* **2008**, *22*, 632.

(2) Kourti, T. Process analysis and abnormal situation detection: from theory to practice. *Control Systems, IEEE 22*, 10.

(3) Westerhuis, J.A.; Kourti, T.; MacGregor, J.F. Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics* **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. 13 (2002) 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* 5 **1995**, 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. Chemometrics,* **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. Chemometrics*, **2001**,*15,* 485.

(11) Wangen, L.; Kowalski, B. A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometrics* **1989,** *3*, 3.

(12) Qin, S.J.; Valle, S.; Piovoso, M.J. On unifying multiblock analysis with application to decentralized process monitoring. *J. Chemometrics* **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. *Chemometr. Intell. Lab.* **2016**, *Accepted for publication*, doi: 10.1016/j.chemolab.2016.03.020

(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, *Desalination and Water Treatment* **2013**, *51*, 6401.

(18) Wagner, J. Membrane Filtration Handbook: Practical Tips and Hints; Minnetonka, MN: Osmonics , 2001.

POSTER I

A. Klimkiewicz, P.P. Mortensen, C.B. Zachariassen, F.W.J. van den Berg

The value of historical data in the optimization of biomanufacturing

Scandinavian Symposium in Chemometrics (SSC13), Stockholm

The value of historical data in the optimization of biomanufacturing

Anna Klimkiewicz[1,2,*], Christian B. Zachariassen[1], Peter Paasch Mortensen[1] and Frans W.J. van den Berg[2]

1) Novozymes A/S, Kalundborg, Denmark;

2) University of Copenhagen, Denmark;

*) akcz@novozymes.com

Aim

At Novozymes granulation processes, drying operations turning enzymes into end-products, are run in a semi-continuous fashion. As a consequence there are no fixed or clear time-relationships between successive unit operations in the granulation process flow. To get a better insight into system dynamics and move towards process optimization/variance reduction the powerful procedure of auto- and cross-correlation on historical data will be used. The main prerequisites for process time series modeling are: 1) sensible extraction of the data from the databases, 2) pre-processing of the data, and 3) interpretation of the findings with engineering intuition. And, as for most data-analytical problems, textbook theory is often only the starting points for the investigation!

Conclusions

The potential of historical full-scale process data and related product quality attributes are explored. The combination of all the variables affecting the system, and their auto- and cross-correlations, carry valuable information for process understanding. Indisputably, most attention needs to be placed on the data preparation and the effects of different signal preprocessing techniques, rather than modeling. And the final results need to be carefully evaluated in cooperation with process operators and engineers. Autocorrelation analysis identified a previously unknown periodic fluctuation in a fluidized bed unit operation. It is anticipated that this periodicity is related to sub-optimal control over the drying process – an opinion supported by the cross-correlations between different process tags.

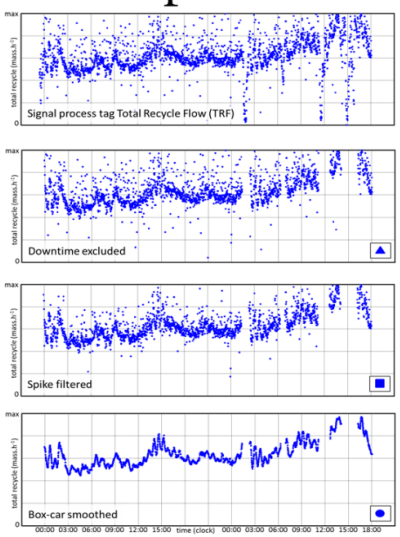
Data clean-up

Historical data - Process data extracted with equidistant sampling form a time series. The raw data include various artifacts due to production breaks, sensor failures, signal saturation, etc. If the aim is to analyze the Normal Operational Conditions (NOC) these situations should be removed.

Downtime excluded - Data irrelevant for the process dynamics was excluded. Artifacts introduced by flow-stoppages are present up till 30 min post-downtime and this adjacent data is also taken out of the analysis.

Spike filter - A window-based filter was applied to remove the most apparent outlying (single) points from the data, judging each measurement by a window-based threshold value.

Smoothing - A box-car/window moving average filter has been applied to reduce the contribution of high-frequency signal components (instrument noise, local turbulent flow behavior, etc.).

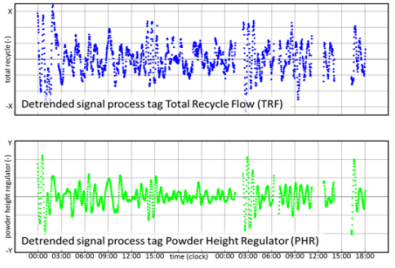


Cross-correlation

If two process signals are shifted or offset in time, simple correlation analysis might be misleading. Cross-correlation functions are more appropriate where the maximum value expresses the properly aligned signals. It is an asymmetric function of two time series, a function of time lag τ:

$$for\ \tau = 0, 1, 2, \ldots \qquad \rho_{xy}(\tau) = \frac{1}{N-\tau}\sum_{t=1}^{N-\tau}\frac{(x(t)-\bar{x})(y(t+\tau)-\bar{y})}{\sqrt{s_x^2 s_y^2}}$$

$$for\ \tau = 0, -1, -2, \ldots \qquad \rho_{xy}(\tau) = \frac{1}{N-\tau}\sum_{t=1}^{N-\tau}\frac{(x(t-\tau)-\bar{x})(y(t)-\bar{y})}{\sqrt{s_x^2 s_y^2}}$$

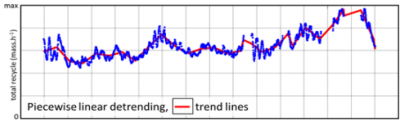


TRF & PHR - The granulate fraction sieved out after a drying bed, not meeting the specification (both too coarse or too fine) is represented by the tag "Total Recycle Flow". Percentage of maximal rotary valve speed controlling the volume/amount of granulate dried in the fluid bed is given by the process tag "Powder Height Regulator". There is a delayed response to a common stimulus that affects both series. The maximum cross-correlation is observed at a lag of minus 6 minutes, which is the case for the data from one single production campaign as well as for data augmented over 10 production campaigns.

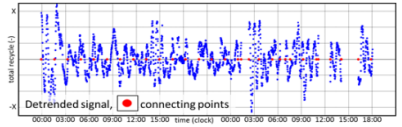
Detrending

The semi-continues granulation process experiences regular set-point changes, resulting in non-stationary time series. Detrending (piecewise linear polynomial fitting, differencing, ARIMA modeling, etc.) is a form of high-pass filtering, which often forms the most critical step in serial correlation analysis.

Piecewise linear detrending - The piecewise-fitted trend lines tracks the lowest frequencies – the large scale variation dictated by the set-point values such as production speed and/or recycling volumes.

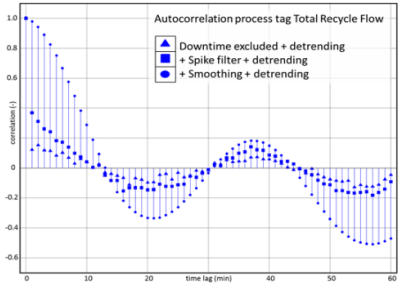


Detrended signal - The residuals from the trend lines have the low frequencies removed. The signal is now comparatively stationary with respect to mean and variance.



Autocorrelation

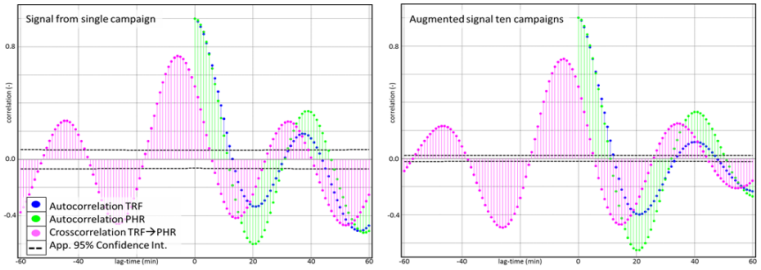
Autocorrelation refers to the correlation of a time series with its own past and future values (process lags, τ). Positive correlations might be considered a specific form of persistence, a tendency for a system to remain in the same state from one observation to the next, while negative correlations indicate feed-back in the system. For large lags noise will be uncorrelated so periodicities are easier to detect from autocorrelograms than from the original process data. The autocorrelation function for time differences is:

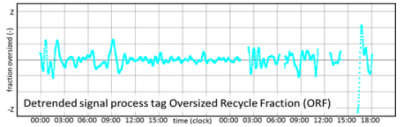


$$\rho_x(\tau) = \frac{1}{N-\tau-1}\sum_{t=1}^{N-\tau}\frac{(x(t)-\bar{x})(x(t+\tau)-\bar{x})}{\sqrt{s_x^2}}$$

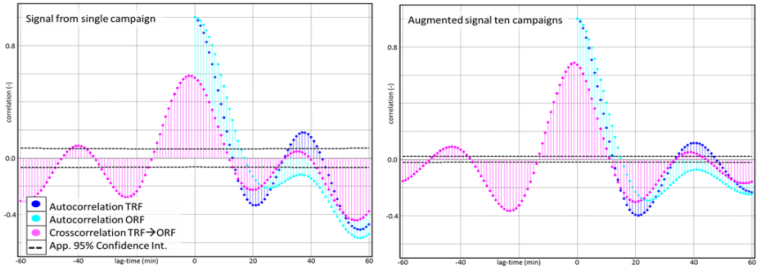
$\bar{x}$: mean of the $N-\tau$ -1 observations
$x(t)$: value at time t
$x(t+\tau)$: value at time $t+\tau$
s_x^2: variance of the $N-1-\tau$ observations

The examined Total Recycle Flow signal showed some unexpected periodicity. Every 20 minutes the recycle flow is adjusted. This could hardly be registered without detrending, marginally for most levels of preprocessing + detrending, but by far the clearest after smoothing + detrending.





TRF & ORF - Granulate particle amount sieved out after the drying bed (due too high diameters) is represented by the tag "Oversized Recycle Fraction", a measurement located upstream compared to the total recycle flow sensor. The cross-correlation reaches its maximum at a lag of minus 1 minute, which is consistent over different and augmented campaigns.



It was not possible to indisputably identify the source of the observed periodicity, but it is suspected to be related to 1) the capacity of the granulation system, 2) mismatch between the location of the feedback signals, 3) delayed response in the control structures and/or 4) external sources with periodicity (e.g. the steam supply system, etc.).

novozymes®
Rethink Tomorrow

POSTER II

A. Klimkiewicz, F.W.J. van den Berg

A chemometric approach to the optimization of bio-industrial processes

Chemometrics in Analytical Chemistry (CAC-2014), Richmond, Virginia

A chemometric approach to the optimization of bio-industrial processes

UNIVERSITY OF COPENHAGEN

Anna Klimkiewicz[1,2,*] and Frans W.J. van den Berg[2]

1) Novozymes A/S, Kalundborg, Denmark

2) University of Copenhagen, Denmark

*) akcz@novozymes.com

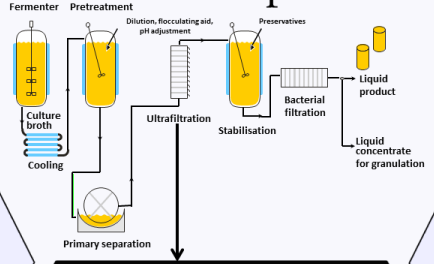
Introduction

Massive amounts of *process tags* are collected with high frequency in biomanufacturing for specific and dedicated univariate monitoring tasks and closed-loop control applications. This produces large amounts of data which are seldom used outside their direct scope. It can however be of great benefit to combine all available data, structure it and extract relevant information from the full set via process chemometrics tools. In this work we highlight the value of historical production data and multivariate statistical data analysis as a starting point of variance reduction and capacity optimization in an industrial-scale continuous enzyme purification process.

Conclusions

The problem of flux decline in ultrafiltration is very complex and multifaceted, even more so in case of a full scale bio-industrial recovery process with numerous sources of variation. The method of connecting historical data from continuous processes to the flux quality parameters is a subject of ongoing research. Process chemometric tools applied on these large amounts of production data offer a good overview during exploratory problem solving, for instance, help optimization engineers to focus attention on suitable target areas or to asses the effect when manipulating process parameters upstream so that no adverse effect downstream is observed.

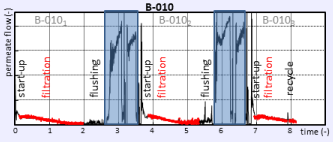
Enzyme recovery, an (almost) continuous process



What's a batch?

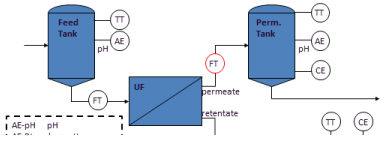
UF *batches* vary considerably in length and volume

- **Operation mode:** discontinuous feed supplied to the membrane plant



- **Unit specific nuances** - i.e. start-up phase finishes when dry matter set-point is reached in the retentate stream leaving the UF
- **Filtration sequences**: separation based on Cleaning-In-Place demands → batch

UF capacity optimization



Query the historian

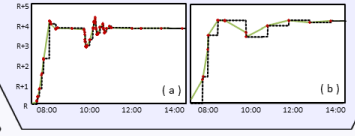
Data can be compressed at any or all of the following steps:

Sensor → Process Control System → Interface Node → Historian

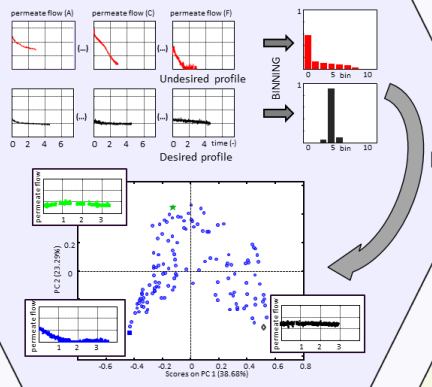
Three most relevant functions for data retrieval:

- **Sampled** (–) interpolated between the two last logged points (equidistant)
- **Archived** (–) return the previous logged value (equidistant)
- **Compressed** (♦) only the true logged points (non-equidistant)

Example - two on-line dry matter measurements and the effect of the dead band settings, 0.2 R (a) and 1.0 R (b)



How to qualify the flux?

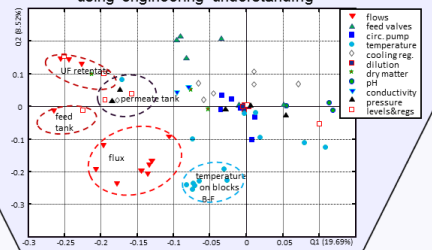


- **Problem**: *UF flux* (permeate volume per unit area of the membrane per unit of time) varies for both the same product and between different products.
- **Aim:** understand *batch-to-batch* variations in the performance of ultrafiltration.

How to select sensors?

Unsupervised methods for batches varying in length:

- Multilevel Simultaneous Component Analysis (MLSCA)
- Common Components and Specific Weights Analysis (CCSWA)
  - → plot similarities between process tags
  - → analysis for clusters and outliers using engineering understanding

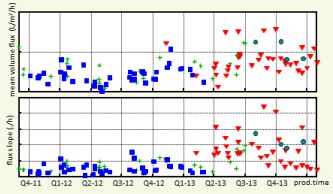


Comparing batches

- Mathematical basis set

Flux as a function of batch production (calendar) time

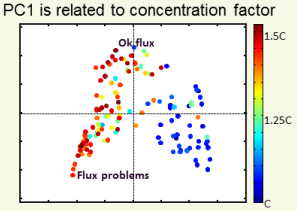
→Higher concentrated group



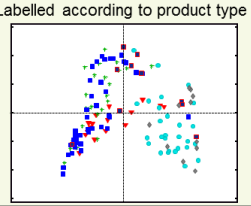
= Introduction of changes in fermentation and recovery SOPs

• Flux quality map

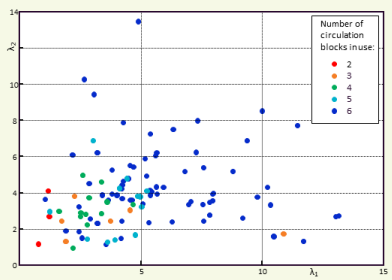
→ PC1 is related to concentration factor



→ Labelled according to product type

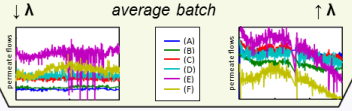


• CCSWA



Saliences = How representative is a batch for the *average batch*

↓ λ ↑ λ



novozymes®

Rethink Tomorrow

DEPARTMENT OF FOOD SCIENCE
FACULTY OF SCIENCE · UNIVERSITY OF COPENHAGEN
PHD THESIS 2016 · ISBN 978-87-7611-994-2

ANNA KLIMKIEWICZ

Multivariate Statistical Process Optimization in the Industrial Production of Enzymes

This thesis focusses on solutions for a more extensive use of full-scale historical production records in data mining, process optimization and problem-solving in the bioindustry. In modern biotech production, a massive number of diverse measurements, with a broad diversity in information content and quality, are stored in data historians. This data is rarely used outside its direct scope due to lack of efficient and suitable procedures for thoughtful data retrieval, evaluation, pre-processing and extraction of the information (modeling). This dissertation work is meant to address the challenges and difficulties related to 'recycling' of historical data from a full-scale manufacturing of industrial enzymes.

Specific chemometric modeling techniques designed for the complex data systems have been examined. These methods maintain the natural structure of the analyzed data by blocking information either in the row (production runs) or column (process parameter types) direction. The complex data structures are decomposed into intuitively interpretable solutions as the important patterns in the data are extracted and visualized. When these patterns are realized and understood, it can lead to a better process understanding in a faster way than traditional mechanistic modeling techniques.

SL grafik