UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Process-water characterisation and quality monitoring in the dairy industry – moving towards replacing potable water

PHD THESIS 2017 · PETER BÆK SKOU

Process-water characterisation and quality monitoring in the dairy industry — moving towards replacing potable water

PhD thesis
by
Peter Bæk Skou

Frederiksberg 2017

Department of Food Science
Faculty of Science
University of Copenhagen

Title
Process-water characterisation and quality monitoring in the dairy industry — moving towards replacing potable water

Submission date
November 2nd 2017

Defence date
December 15th 2017

Supervisors
Associate Professor Frans van den Berg
Department of Food Science, Faculty of Science
University of Copenhagen, Denmark

Professor Søren Balling Engelsen
Department of Food Science, Faculty of Science
University of Copenhagen, Denmark

Opponents
Professor Lilia Ahrné (chair)
Department of Food Science, Faculty of Science
University of Copenhagen, Denmark

Associate Professor Wouter Saeys
Department of Biosystems, Science, Engineering and Technology Group
KU Leuven, Belgium

Chief Science Officer, Frank Westad
CAMO Software AS, Norway

Cover illustration from colourbox

ISBN: 978-87-7209-080-1
Printed by SL grafik, Frederiksberg, Denmark (slgrafik.dk)

Table of Contents

Preface

This thesis is submitted to obtain the PhD degree from the PhD school of Science, University of Copenhagen. The PhD project was conducted as part of the Reuse of WAter in the food and bioprocessing inDustry (REWARD) project funded by the Danish council for Strategic Research, Programme Commision on Health Food and Welfare and REWARD partners, including University of Copenhagen, the Technical University of Munich, the Technical University of Denmark, NIRAS, DHI, Novozymes, LiqTech International, TetraPak and Arla Foods Ingredients.

Acknowledgement goes to my principal supervisor, Associate Professor Frans van den Berg, and co-supervisor Professor Søren Balling Engelsen, University of Copenhagen. A special appreciation goes to Frans for always taking time for questions, thorough review of the work done and vast knowledge of general engineering and mathematics as well as specific knowledge on the processes covered in this project.

Arla Foods Ingredients have played a crucial role in letting me sample the processing plants that fit me and my research interests and are acknowledged for this. Hans Henrik Holst and especially Dorrit Thaysen are thanked for being positive and constructive in their collaboration. Many process operators must be thanked for letting us interrupt their daily (and nightly) work, always up for the challenge of finding the right sample valve or climbing a storage tank to collect our samples — none mentioned, none forgotten. In this regard I must also thank all the people who have helped me collect samples at Arla Foods Ingredients (day and night) at several occasions. Thanks goes to Philip Junker Andersen, Iuliana Stoica, Stina D. Aunsbjerg, Thilo A. Berg and Frans van den Berg.

Programme Manager Stephen E. Holroyd, Fonterra is thanked for hosting me at Fonterra's Research and development centre in Palmerston North, New Zealand. Not only was the stay rewarding scientifically, but also personally as my family had the opportunity to join me in New Zealand during the five month stay.

I have enjoyed working in the section for Chemometrics and Analytical Technology with the eventful Christmas lunches, summer trips and overall good atmosphere.

Lastly I thank my family for their support during the PhD study. Especially my girlfriend Lise and our son Alfred for providing the focus to get home in time. I love you both.

Peter Bæk Skou
Frederiksberg, November 2017

List of publications

Paper I

Peter B. Skou, Bekzod Khakimov, Thomas H. Hansen, Stina D. Aunsbjerg, Susanne Knøchel, Dorrit Thaysen and Frans van den Berg (2017). Chemical characterization by gas chromatography - mass spectrometry and inductively coupled plasma - optical emission spectroscopy of membrane permeates from an industrial dairy ingredient production used as process-water, *Journal of Dairy Science*, in press.

Paper II

Peter B. Skou, Thilo A. Berg, Stina D. Aunsbjerg, Dorrit Thaysen, Morten A. Rasmussen and Frans van den Berg (2017). Monitoring Process Water Quality Using Near Infrared Spectroscopy and Partial Least Squares Regression with Prediction Uncertainty Estimation. *Applied Spectroscopy 71*, 410–421.

Paper III

Peter B. Skou, Stephen E. Holroyd and Frans van den Berg (2017). Tutorial – Applying extreme value theory to characterize food-processing systems. *Journal of Chemometrics*, e2896.

Paper IV

Lea G. Johnsen, Peter B. Skou, Bekzod Khakimov and Rasmus Bro (2017). Gas chromatography – mass spectrometry data processing made easy. *Journal of Chromatography A*, 57–64.

Additional papers not included in the thesis

Paper V

Rikke K. Lauridsen, Peter B. Skou, Thomas Rindzevicius, Kaiyu Wi, Søren Molin, Søren B. Engelsen, Kim G. Nielsen, Helle K. Johansen and Anja Boisen (2017). SERS spectroscopy for detection of hydrogen cyanide in breath from children colonised with *P. aeruginosa*. *Analytical Methods*, Advance Article.

Abbreviations and notation

Abbreviations

AFI	Arla Foods Ingredients
AOC	Abnormal operating conditions
BM	Block maxima
BOD	Biological oxygen demand
CIP	Cleaning in place
COD	Chemical oxygen demand
DOC	Dissolved organic carbon
EEM	Excitation emission
EVT	Extreme value theory
FO	Forward osmosis
GC-MS	Gas chromatography mass spectrometry
GEV	Generalised extreme value
ICP-OES	Inductively coupled plasma optical emission spectroscopy
IID	Independent and identically distributed
MIS	Method of independent storms
MSEC	Mean square error of calibration
MWCO	Molecular weight cut-off
NAS	Net analyte signal
NF	Nano filtration
NIR	Near infrared
NIRS	Near infrared spectroscopy
NOC	Normal operating conditions
PARADISe	PARAFAC2 based deconvolution and identification system
PARAFAC	Parallel factor analysis
PLS	Partial least squares
POT	Point over threshold
REWARD	Reuse of water in the food and bioprocessing industry
RO	Reverse osmosis
SBSE	Stir-bar sorptive extraction
SPME	Solid-phase microextraction
TC	Total carbon
TOC	Total organic carbon
UF	Ultra filtration
UV	Ultraviolet
WPC	Whey protein concentrate

Algebra notation

$\underline{\boldsymbol{X}}$	Underline, uppercase, italics and bold: data cube
$\boldsymbol{X}$	Uppercase, italics and bold: matrix
$\boldsymbol{x}$	Lowercase, italics and bold: column vector
$\boldsymbol{X}^T$, $\mathbf{x}^T$	Transpose of a matrix or vector
X, x	Italics and non-bold: scalar

Abstract

The dairy industry is a major consumer of potable water, with large volumes being used in cleaning operations and for facility needs such as heating / cooling media. In an attempt to bring down water use and also minimize discharge of wastewater, process-water is explored as an alternative source. Process-water stems from processing equipment usually already present in the dairy industry and mainly concerns membrane filtration permeate as well as evaporator condensate. During processing process-water is reclaimed and, if suitable, it can replace potable water where huge savings both on monetary and environmental impact can be achieved. Since process-water is reclaimed from a production process, and all processes are subject to variability, the quality must be guaranteed at all times to ensure safe use. Ideally this is done via real-time on-line measurement systems.

The aim of this thesis is to perform detailed chemical characterisation of process-water streams and to identify and test relevant measurement techniques, thus establishing basic knowledge on process-water quality monitoring possibilities.

Two process-water streams at the Arla Foods Ingredient's processing facilities have been investigated in the thesis work: 1, membrane permeate and 2, evaporator condensate.

In membrane permeate urea was found as the main organic compound permeating the membrane, but also relatively large molecules were also found to permeate in low concentrations. Near infrared spectroscopy was investigated as a potential monitoring technique and was found to be sufficiently sensitive in a laboratory set-up. However, the signal-to-noise ratio was so low that thorough uncertainty estimation was needed to ensure confidence in the predictions.

Evaporator condensates were characterised with a number of techniques and the results indicate that aromatic amino acids consistently were present in condensates from several processing lines. However, no conclusive identification was reached pointing towards the necessity to employ other analytical techniques that focus on volatile compounds.

The investigated processes were found to be very stable and consequently left collection of samples with high organic load up to chance. This led us to explore extreme value theory as a way to characterise production processes' distribution and dynamics.

In conclusion the investigated process-water streams appeared to be very clean, making quantification and identification of possible contaminating compounds challenging.

Resumé

Mejeriindustrien er en storforbruger af drikkevand, hvor store mængder bliver brugt i rengøringsprocesser og til anlægsdrift såsom varme / kølevand. I et forsøg på at nedbringe vandforbruget og samtidigt minimere udledningen af spildevand udforskes procesvand som en alternativ kilde. Procesvand stammer fra procesudstyr som ofte allerede er tilstede i mejeriindustrien og primært vedrører membranfiltreringspermeat og fordamperkondensat. Under processering genvindes procesvand og, hvis egnet, kan det erstatte drikkevand, hvorved store besparelser både monetært og miljømæssigt kan opnås. Eftersom procesvand genvindes fra en produktionsproces, og alle processer er genstand for variabilitet, skal kvaliteten garanteres til hver en tid for at sørge for sikker brug. Ideelt gøres dette via målesystemer direkte tilkoblet processen i realtid.

Målet med denne afhandling er at udføre detaljeret kemisk karakterisering af procesvandstrømme samt identificere og teste relevante målemetoder for således at lægge grunden for monitoreringsmuligheder af procesvandskvalitets.

To procesvandsstrømme hos Arla Foods Ingredients produktionsfaciliteter er blevet undersøgt i denne afhandling: 1, membranpermeate og 2, fordamperkondensat.

I membranpermeatet blev urea fundet som den vigtigste organiske forbindelse, som gennemtrængte membranerne, mens relativt store molekyler også blev fundet i lave koncentrationer. Nærinfrarødspektroskopi blev undersøgt som en potentiel måleteknik and blev vurderet til at være tilstrækkelig følsom i laboratorie opstilling. I midlertidigt var signal-til-støj forholdet så lavt, at omhyggelige usikkerhedsestimater var nødvendige for at sikre tillid til prædiktionerne.

Fordamperkondensat blev karakteriseret med en mængde teknikker og resultaterne indikerer, at aromatiske aminosyrer konsekvent var tilstede i kondensat fra forskellige proceslinjer. Dog blev ingen endegyldig identifikation nået, hvilket peger mod nødvendigheden af at køre andre analytiske teknikker som fokuserer på flygtige forbindelser.

De undersøgte processer blev vurderet som meget stabile, og derfor blev indsamlingen af prøver med højt organisk indhold efterladt til tilfældigheder. Dette førte til at undersøge ekstremværditeori som en måde at karakterisere produktionsprocessers fordeling og dynamik på.

Afslutningsvist viste det sig at procesvand lader til at være meget rent, så rent at kvantificering og identifikation af forurenende forbindelser var meget udfordrende.

Chapter 1

Introduction

Water scarcity has been a main topic for the United Nations and the World Health Organization (WHO) for many years.[1,2] Despite these efforts WHO expects that by 2025 half the world's population will be living in water-stressed areas. Denmark as a whole is not a water-stressed country,[3] regardless some areas are running out of potable water limiting production capacity. It is therefore of great interest to minimize potable water consumption in Denmark.

The food industry and particularly the dairy industry is a major consumer of water with an estimated usage of 1–5 litres of water per kg of milk processed.[4] Data on water usage and wastewater discharge in various industries from Denmark in 2015 is reported in Table 1.1. In total all industrial manufacturing consumed 55 million cubic meters of water with the food and beverage industry as the largest contributor. Breaking down the water use in the food and beverage industry into the various sectors, the dairy and meat industry are the largest consumers. Many food producers will have their own water sources which lowers the expense for water use. However, wastewater discharge costs must be settled with municipalities, and as seen in Table 1.1 expense are larger than for consumption.

These numbers should be seen in light of the total water consumption in Denmark of 800 mio. m^3, where agriculture and horticulture account for almost 200 mio. m^3.[5] This is the general trend in most of the world, but a collective effort should still be implemented.[6]

In the dairy industry water is mostly used for cleaning of processing equipment (via cleaning in place, CIP), pasteurisation (water for indirect heating e.g. in plate heat exchangers), pump sealing in centrifugal pumps, production of steam and as a cooling medium.

Milk and whey consist mainly of water and in cheese manufacturing and whey processing this excess water is removed to concentrate the valuable mass / product (i.e. protein, lactose and fat). Traditionally this water has been discharged as wastewater, but if the quality is sufficiently heightened to replace potable water

Table 1.1: Water consumption (mio. m^3 & mio. DKK) for all industries consuming >1 mio. m^3 in Denmark in 2015.[5] Food and beverage industry is further divided into sectors.

	Water consumption		**Wastewater discharge**	
	mio. m3	mio. DKK	mio. m3	mio. DKK
Food & beverage	**27.7**	**258.5**	**23.9**	**631.1**
Dairy	7.7	54.0	4.6	122.5
Meat	7.4	71.2	6.8	145.8
Other food	4.8	40.8	2.4	199.7
Beverage	3.0	16.2	1.3	17.6
Fishery	2.5	47.7	6.8	105.7
Baking	2.3	28.2	2.1	39.0
Chemical	**7.6**	**85.9**	**5.3**	**99.2**
Plastic, glass & concrete	**5.7**	**22.4**	**4.3**	**38.4**
Medical	**4.4**	**58.0**	**4.0**	**60.8**
Oil refinery	**4.1**	**13.0**	**3.9**	**50.1**
Metal	**2.0**	**13.6**	**2.0**	**26.4**
Engineering	**1.3**	**15.1**	**1.3**	**31.1**

there is a huge potential to reduce intake of potable water and minimize discharge, hence an environmental and economic advantage on two fronts.

In the present case, Arla Foods Ingredients (AFI) receive whey either *as is* from nearby dairies or in concentrated form from dairies further away. In the latter case reverse osmosis (RO) membrane filtration is used to reduce the transportation needs and costs by concentrating the whey five times. Regular cheese whey contains approximately 0.8% protein, 0.05% fat, 5% lactose, 0.7% minerals and 94% water.[7] During weigh-in at AFI's production facility each tank truck's whey is analysed for protein, fat and lactose content. The collected whey is then combined in large balance tanks (150–250 m^3). In general (not AFI specific but based on Walstra et al.[7]), the processing starts with ultra filtration (UF) where the protein fraction is collected. Dry matter content is increased as much as possible using UF in dilution mode (previously known as diafiltration[8]) where the retentate is diluted with water and then processed again by UF to remove low molecular weight compounds.

The protein fraction from the UF process is further concentrated by evaporation until it is finally spray-dried into the finished product. The products that are produced include, but are not limited to: whey protein concentrate, whey protein isolate (depending on the degree of dilution mode used) as well as more specific protein fractions or even specific proteins. Specific proteins and protein fragments can be obtained by additional fractionation / isolation by chromatographic columns and fragments by hydrolysis. The use of these products range from alternative (low commodity) protein sources, functional ingredients (e.g. sta-

bilising foams and emulsions in processed foods), ingredients to create novel food products, and to ingredients designed to fulfil specific nutritional requirements.

In the UF processing of whey only the collected part or *retentate* was dealt with. Everything that is not retained in a filtration process passes through or *permeates*. UF whey permeate contains mainly lactose and minerals. By utilising RO membranes to filtrate the UF permeate, the lactose and minerals can be harvested, and similar to the protein fraction, this fraction is further concentrated and eventually also spray dried.

As explained above, AFI extract the valuable material from cheese whey leaving behind water. I define this type of water *process-water* as a consequence of two properties: (a) the source, or raw material, is of a much higher quality compared to e.g. wastewater due to the fact that it is derived from a food and food production, and (b), due to the processing technologies used to extract it.

The desire of the food production industry is to use such process-water streams to replace potable water instead of discharging it — figuratively going from Figure 1.1a to b. The ultimate goal is to eliminate intake of potable water and only discharge clean water without compromising production capacity, quality or safety.

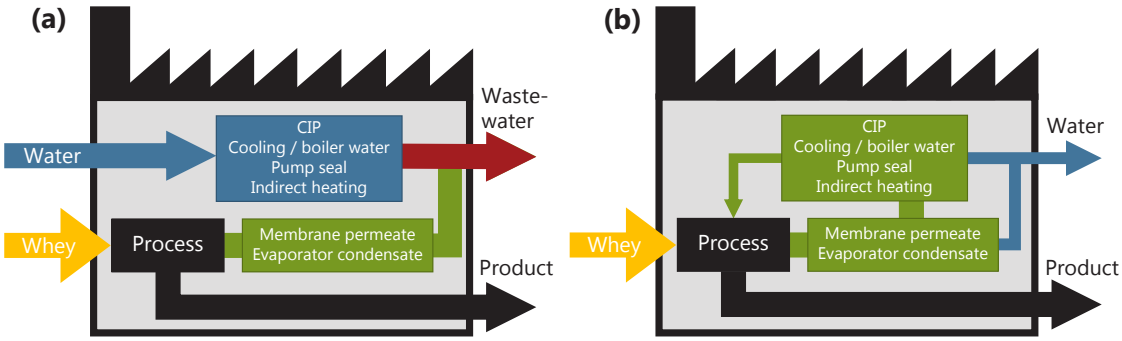

Figure 1.1: Illustration of the principle of process-water in the dairy industry from no use (a) to complete use (b).

One of the most promising processing technologies for recovering process-water within dairy processing is membrane-filtration, especially *reverse-osmosis* (RO) membrane filtration. The motivations are that RO filtration is already used extensively in the dairy industry and is expected to increase,[9,10] and it is already widely used to purify sea-water and wastewater.[11,12] Additionally, evaporator condensate has been regarded as a very clean water stream. For these reasons AFI's processes are obvious places to investigate the potential for further using process-water.

1.1 Outline

The thesis is built up around the different studies performed during the project. In order to avoid repetition the published papers serve as integrated parts of the thesis and very little is reiterated in the additional text. Furthermore, theory on the established techniques used is kept at a minimum and the reader is referred to other works for detailed explanation when relevant.

Chapter 2 concerns detailed chemical, and to a lesser extent microbial, characterisation of process-water reclaimed from membrane filtration processes. The chapter opens with a brief description of how membranes are composed and how separation takes place after which studies from literature are discussed and opportunities for research are identified. **Paper I** presents a study on samples from an RO process at AFI with derivatisation based gas chromatography coupled with mass spectrometry (GC-MS) and inductively coupled plasma optical emission spectroscopy (ICP-OES).

Chapter 3 covers quality monitoring strategies for membrane permeate. The main organic compound found in the characterisation, urea, is followed with near infrared spectroscopy (NIRS) and partial least squares regression (PLS), a study formalized in **Paper II**. The chapter continues with some comments to **Paper II** and concludes with an attempt to establish a NIRS based tensor calibration.

Chapter 4 moves the focus to another process-water stream, namely evaporator condensate from AFI. Two investigations are presented which try to identify which organic compounds that are present in this process-water stream.

Chapter 5 presents the concept of extreme value theory (EVT) and how to apply this in food production processes, exemplified by a case study in **Paper III**.

Chapter 6 concludes on the results obtained and gives perspectives for future research concerning the use of reclaimed process-water in the dairy industry.

Chapter 2

Characterisation of membrane permeates

In full scale spirally-wound membrane filtration processes the feed stream enters the membrane at the perimeter of the element under pressure in a crossflow fashion as shown in Figure 2.1. Permeate flows to the center and escapes via a transport tube. The membrane unit consists of flat sheets of the membrane material that are spiral-wound around this center tube. An example of a commercial membrane element is given in Box 1. Several membrane elements are combined in one casing and several casings are linked in series as well as parallel to increase the processing surface area. The casings, together with the pumps, feed / balance tanks and measurement-and-control facilities form the membrane unit, and an example of a pilot or small scale production plant can be seen in Figure 2.2.

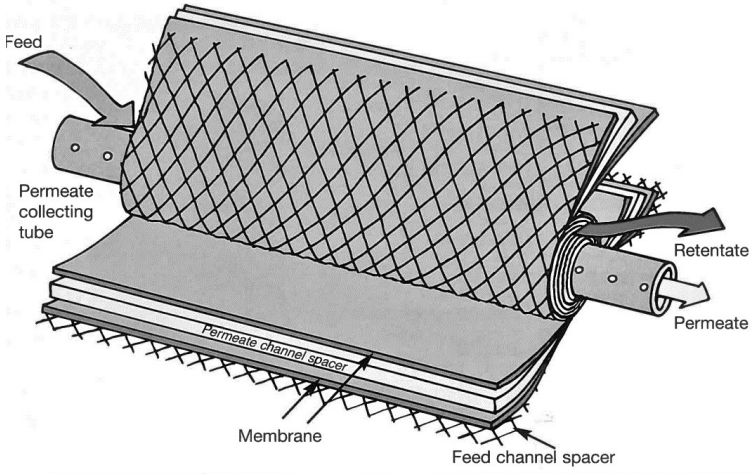

Figure 2.1: Illustration of a membrane unit from Nielsen.[13]

Membrane technologies are often differentiated based on their molecular weight cut-off (MWCO) value. The MWCO is defined as 90 % rejection of molecules with this weight or greater. The membrane type with the highest MWCO — meaning letting small molecules pass — is micro-filtration followed by UF, nano-filtration (NF) and finally RO membranes. However, describing a RO membrane

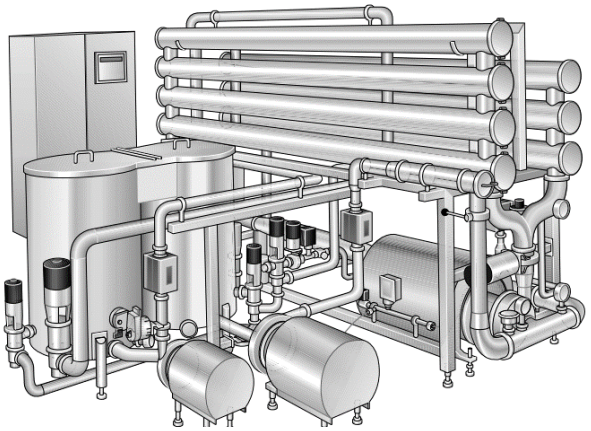

Figure 2.2: Illustration of a membrane processing plant from Bylund.[14]

(or any other membrane technology) by its MWCO can give a false impression that molecules are excluded based on size alone. RO membranes separates dissolved solutes from water by what can best be described as a solution-diffusion mechanism.[11] Instead of physically blocking dissolved molecules from passing the membrane wall based on size and weight, the permeating molecules dissolve in the membrane material and pass through. Size does have an effect on diffusion rates meaning that large molecules will travel very slowly through the membrane material.

Ozaki and Li[15] investigated the rejection of organic compounds by ultra-low pressure RO filtration and found that urea and acetic acid (which have very similar molecular weight of 60.06 and 60.05 g·mol^{-1}, respectively) were rejected with equal efficiency (30 %) at pH 3. By changing the pH (Figure 2.3) the rejection of urea remained nearly constant while rejection of acetic acid increased with increased pH. Acetic acid has a pKa of 4.76 meaning that at pH 3 it will be mostly protonated and at pH 9 fully dissociated with a resulting negative charge. Urea is not able to dissociate in this pH range and remains neutral. The authors speculated that the negative charge on the dissociated acetic acid causes electrostatic repulsion to the membrane material, thereby rejecting transport over the wall, while urea remains neutral. Lee and Lueptow[16] also looked into the rejection of urea by RO membranes and speculated that urea is poorly rejected due to a high affinity towards the membrane material. Membrane element producers also use the option to modify (*spike*) the basic chemistry to improve separation mechanisms (e.g. enhancing the membrane hydrophilic or hydrophobic properties) and to counteract surface and internal fouling.

Box 1: RO membranes, an example

DOW Hypershell RO-8038 is a FDA approved RO membrane with a composite membrane barrier consisting of a thin-film aromatic polyamide barrier, polysulfone interlayer and polyester support web with a polypropylene outer shell. The specifications are shown in Table 2.1.

Table 2.1: Data sheet for DOW Hypershell RO-8038.

Dimensions	**RO-8038**
Length	965 mm
Diameter	200 mm
Active area	34.4 m^2
Feed spacer	0.84 mm
Operating and Cleaning limits	
Max. operating pressure	54.8 bar
Max. recirculation cross-flow	18.2 $m^2 \cdot h^{-1}$
Free chlorine tolerance	None
pH range	2–11
Max. operating temperature	
pH 2–10	50°C
pH > 10	35°C
Hydrogen Peroxide usage limits	
Continuous operation	20 ppm
Short-term cleaning (max 20°C)	1000 ppm

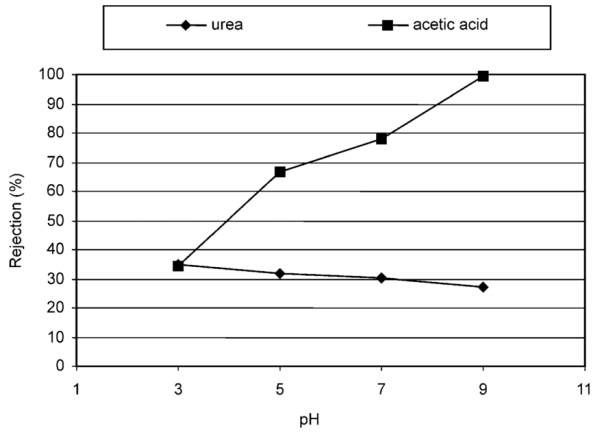

Figure 2.3: Rejection efficiency of urea and acetic acid by a low-pressure RO membrane at different pH; reproduced from Ozaki and Li.[15]

Theoretically any compound can permeate (in infinitely small amounts) a RO membrane. However, in practice molecules must be able to sufficiently dissolve into the membrane and be small enough to diffuse through during the time-span of a process run.

Membrane processing has become an important technology for purifying water.[11] One of the most extreme cases is found in Western Australia, which has experienced increasing water stress as a result of reduced rainwater precipitation. To counter this effect wastewater has been treated with UF and RO membrane filtration and subsequently ultraviolet (UV) radiation and is currently being considered as an alternative potable water source.[12] Since wastewater can originate from many sources and the intended use is direct consumption 375 chemicals were identified and screened to characterise the dissolved organic carbon (DOC, see Box 2) and ensure safe use. It was concluded that none of the contaminants permeating posed a health risk. Interestingly the detected chemicals could not account for the total DOC, implying that despite very broad compound coverage not all carbon could be accounted for.

Box 2: Bulk water quality parameters

Traditionally (drinking) water quality has been defined from bulk water quality parameters including but not limited to: biological oxygen demand (BOD), chemical oxygen demand (COD) and total organic carbon (TOC). In the following text a brief overview of the different analyses is given. For extensive discussions and instructions see Standard Methods For The Examination of Water and Wastewater.[17]

BOD is an empirical test to determine the oxygen requirements of wastewater. Samples are inoculated with aerobic organisms (preferably from the biological treatment plant processing the wastewater) that break down the organic material under oxygen consumption, typically determined over five days at 20°C. Whether the micro-organisms are able to breakdown all the organic material cannot be known rendering the method mostly useful for only biological treatment plants and less useful for characterisation of water for other purposes.

COD is defined as the amount of oxidant used to oxidise the sample. Usually dichromate ($Cr_2O_7^{2-}$) is used as oxidant and the consumed amount of oxidant is determined by excess-titration with ferrous ammonium sulphate. Since dichromate is not a strong oxidant not all organic material will be oxidized while inorganic compounds may oxidise. The advantage is that the method is relatively inexpensive and fast to perform.

Box 2: Bulk water quality parameters

TOC is determined by total oxidation of carbon to carbon dioxide gas by e.g. a high temperature or UV radiation. Carbon dioxide gas is measured with an infrared analyser. The total oxidation and specific quantification of carbon gives a clear picture of the organic load in the water sample and the infrared gas phase detection makes the method very sensitive. Unfortunately due to the strong oxidation and sensitive detection dedicated equipment is needed for TOC determination.

Samples can be pre-treated before carbon quantification leading to different fractions and definitions: total carbon (TC; analysing sample as is), total inorganic carbon (TIC; acidifying and purging sample without oxidation leading to quantification of mainly carbonates), dissolved organic carbon (DOC; quantification after filtering sample over 0.45 µm filter). TOC is the different between TC and TIC.

Extensive research has been done on purifying wastewater, model solutions of wastewater, whey, milk and evaporator condensates from the dairy industry, as shown in Table 2.2. Studies dealing with UF membranes only have been omitted from the overview since this is more relevant for fractionating of proteins than for production of pure water.

The cause for the extensive list of references can probably be ascribed to the wide application field membranes have already had in the dairy industry;[10] and thus know-how on running these or similar processes is already present in the dairy industry. One of the first works done on a dairy derived process-water stream was performed by Chmiel et al. in 2000.[19] In this study evaporator condensate from a milk concentration process was filtered over a two-stage pilot scale NF membrane plant running for three months at a milk processing company. The COD levels did not exceed 10 $mg \cdot l^{-1}$ in the 28 measurements spaced out across the demonstration run.

In 2008 Vourch et al.[31] did a detailed chemical characterisation of evaporator condensates and one- or two-stage pilot scale RO filtered dairy wastewater. In addition to the regular bulk parameters such as COD, DOC, TOC and conductivity also minerals were quantified, and in the case of evaporator condensate also ethanol, acetone and acetoin by headspace GC coupled with a flame ionization detector.[45] For the RO treated wastewater the carbon contribution of lactose in terms of TOC equivalents was calculated and compared to the actual obtained TOC values. It was concluded that lactose accounted for 76–100% of the organic carbon in the permeate.

More recently Suárez et al.[41–43] investigated the possibility of reusing RO treated ultra high temperature flash-cooler condensate. A pilot scale RO membrane plant filtered the condensate and evaluated filtration efficiencies based on

Table 2.2: Chronological overview of publications concerning membrane filtration with the aim of purifying dairy wastewater, whey, milk or evaporator condensate, either for the purpose of discharge or reclamation of water. **Membrane type** specifies the primary membrane technology, but other pre-treatments (e.g. coarse filters) or advanced technologies (e.g. rotating disk RO) may have been included. **Water source** specifies the feed for the membranes. All water sources are taken from industrial dairy operations unless *model* is stated, defining that water was prepared e.g. by diluting milk or mixing various ingredients in clean water. **Scale** of operation specifies the size of the filtration operation where lab scale indicates the use of non-industrial sized membrane elements (typically cut-outs of full size membrane elements) and pilot scale uses a small number of full size membrane elements.

Membrane type(s)	**Water source**	**Scale**	**Year Published**	**Refs.**
MF+UF	Wastewater	Pilot	1996	[18]
NF+NF	Evaporator condensate	Pilot	2000	[19,20]
NF, RO+RO	Wastewater	Pilot	2000	[21]
NF+RO	Wastewater model	Lab	2002	[22]
NF	Whey	Pilot	2003	[23]
NF, RO	Wastewater	Pilot	2004	[24]
NF, RO	Wastewater model	Lab	2004	[25]
NF, RO, NF+RO, RO+RO	Wastewater model	Pilot	2005	[26]
NF	Whey	Pilot	2006	[27]
NF, RO	Wastewater	Lab+pilot	2006	[28]
NF	Wastewater model	Lab	2006	[29]
UF, NF, RO, NF+RO	Whey	Pilot	2008	[30]
RO, RO+RO	Wastewater, evap. cond.	Pilot	2008	[31]
RO	Wastewater model	Lab	2008	[32]
NF	Industrial & model wastewater	Lab	2010–2011	[33,34]
UF+NF	Wastewater model	Lab	2011	[35]
UF+NF	Wastewater	Lab	2012	[36]
NF+RO	Wastewater	Lab	2012	[37]
NF	Flash-cooler condensate	Pilot	2013	[38]
FO, RO	Whey	Lab	2013	[39]
RO	Whey model	Lab	2014	[40]
RO	Flash-cooler condensate	Pilot	2014–2015	[41–43]
NF	Wastewater	Lab	2015	[44]

conductivity and COD. In trying to find appropriate long-term operation parameters, up to 100 hour filtration runs were performed. An example of such a run can be seen in Figure 2.4.

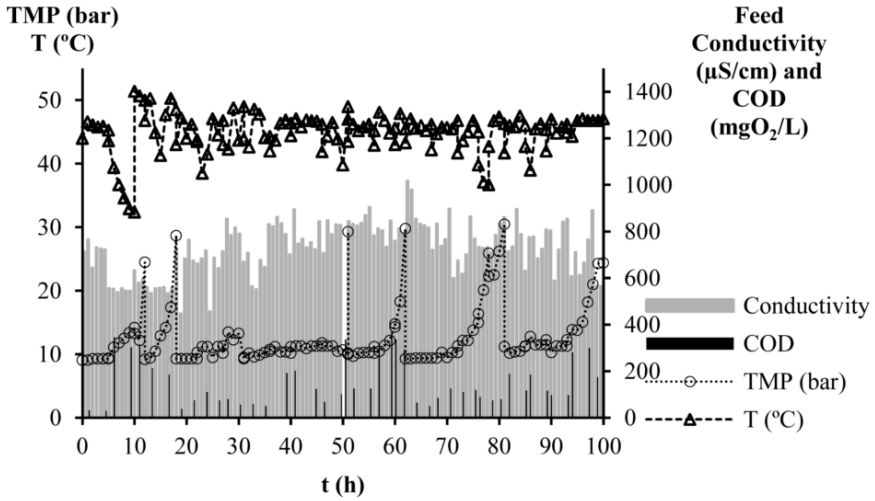

Figure 2.4: Example of permeate quality from a long-term operation of a pilot scale RO membrane filtration; from Suarez et al.[41]

Chmiel et al.[19] demonstrated long-term performance of NF membranes over three months of operation, but did not include the short-time dynamic behaviour. This technology is now outdated as RO membranes succeeded NF for water treatment. The dynamic short-term behaviour of RO permeate quality was elucidated by Suarez et al.[41–43] where it should be noted that the water originates from a different feed stream (flash-cooler condensate). While Vourch et al.[31] characterised the permeate composition in more detail than other studies, no un-targeted screening approach was performed. Un-targeted chemical characterisation of RO has been done in a pilot scale desalination of sea-water[46], but no studies have been found on dairy related processes. Despite the filtration plant in Vourch et al.'s study[31] being at pilot scale size they do not perform long-term operations as done in e.g. Chmiel et al.,[19] thus lowering the transferability to industrial settings.

Looking through the studies in Table 2.2 it is apparent that no extensive studies have been reported on permeate from *industrial* membrane filtration plants. This is problematic for two reasons. Laboratory or pilot scale units seldom run more than a few hours or weeks. Membrane surfaces will foul over time with the fouling layer becoming part of the barrier.[47] Hence, it is questionable how representative theses (academic) studies are for daily industrial practice. Furthermore, the study by Vourch et al.[31] contains the most detailed chemical characterisation of a dairy

derived RO permeate. But, continuing to use selective (or targeted) methods increases the risk of overlooking potential important compounds, especially potential carry-overs from the complex formulations of CIP media in industry.

In the following paper we suggest un-targeted and less biased analytical methods to characterise the chemical compounds permeating an industrial scale membrane filtration process at a dairy ingredient producer.

Paper I

Chemical characterization by gas chromatography - mass spectrometry and inductively coupled plasma - optical emission spectroscopy of membrane permeates from an industrial dairy ingredient production used as process-water

Peter B. Skou, Bekzod Khakimov, Thomas H. Hansen, Stina D. Aunsbjerg, Susanne Knøchel, Dorrit Thaysen and Frans van den Berg

Journal of Dairy Science, in press

J. Dairy Sci. 101:1–12
https://doi.org/10.3168/jds.2017-12950

Chemical characterization by gas chromatography-mass spectrometry and inductively coupled plasma-optical emission spectroscopy of membrane permeates from an industrial dairy ingredient production used as process water

Peter B. Skou,*[1] Bekzod Khakimov,* Thomas H. Hansen,† Stina D. Aunsbjerg,* Susanne Knøchel,* Dorrit Thaysen,‡ and Frans van den Berg*
*Department of Food Science, and
†Department of Plant and Environmental Sciences, Faculty of Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg C, Denmark
‡Arla Foods Ingredients, Sønderupvej 26, DK-6920 Videbæk, Denmark

ABSTRACT

Reusing reverse osmosis (RO) membrane permeate instead of potable water in the dairy industry is a very appealing tactic. However, to ensure safe use, the quality of reclaimed water must be guaranteed. To do this, qualitative and quantitative information about which compounds permeate the membranes must be established. In the present study, we provide a detailed characterization of ultrafiltration, RO, and RO polisher (ROP) permeate with regard to organic and inorganic compounds. Results indicate that smaller molecules and elements (such as phosphate, but mainly urea and boron) pass the membrane, and a small set of larger molecules (long-chain fatty acids, glycerol-phosphate, and glutamic acid) are found as well, though in minute concentrations (<0.2 μM). Growth experiments with 2 urease-positive microorganisms, isolated from RO permeate, showed that the nutrient content in the ROP permeate supports limited growth of 1 of the 2 isolates, indicating that the ROP permeate may not be guaranteed to be stable during protracted storage.
Key words: dairy ingredient production, process water, membrane filtration, quality characterization

INTRODUCTION

Reducing the overall water footprint has become an important objective in the dairy processing industry due to increasing cost of discharge and intake as well as limited availability of potable water. The potable water consumption can be greatly reduced through efficient use of process waters, such as membrane permeate and evaporator condensate, in different areas of the production including cleaning processes.

In dairy processing facilities, membrane filtration technologies are already frequently used to create new products. These technologies can be characterized by their ability to separate molecules and constituents of different sizes. At Arla Foods Ingredients, Viby, Denmark, whey, a by-product from cheese production facilities, is processed into new products. First, UF membranes are used to retain whey proteins, and the permeate is processed through a 2-stage reverse osmosis (**RO** plus RO polisher; **ROP**) membrane plant to collect lactose. The collected whey proteins are used downstream to produce various products and constituents, whereas the lactose is used as a food ingredient (e.g., in infant formula). The focus of this investigation is the use potential of the ROP permeate as a process water stream.

Although use of process water is an attractive option, microbiological safety is of major importance in the food and dairy industry and must be ensured at all times. To do this, real-time online monitoring of relevant parameters should be considered (Casani et al., 2005). Process analytical technology provides the principles for real-time online monitoring (van den Berg et al., 2013), but to select the appropriate measurement technique the target molecule(s) must first be identified. The RO(P) permeates from dairy-derived water have been reported to have a very low organic load and low conductivity indicating minimal concentrations of organic compounds and minerals. Organic load has traditionally been expressed by classical, cumulative numbers such as total organic carbon, chemical oxygen demand, and total nitrogen (Vourch et al., 2005, 2008). However, these cumulative measurements do not provide insight into which specific compounds permeate the membranes. Vourch et al. (2008) looked at selected

Received March 29, 2017.
Accepted August 24, 2017.
[1]Corresponding author: peter.b.skou@food.ku.dk

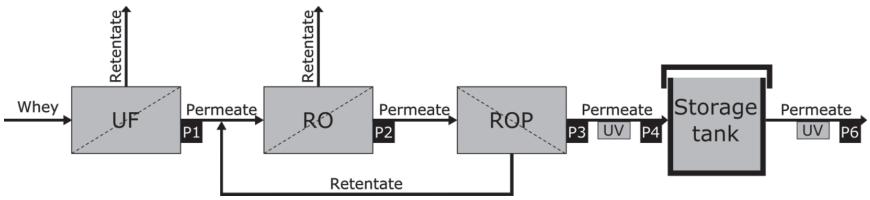

Figure 1. Schematic overview of UF processing and reverse osmosis (RO) plus RO polisher (ROP) lactose and water recovery plus storage. The process water sampling points are marked P1 to P6.

organic compounds and ions in membrane permeates and vapor condensates. A different approach is taken (e.g., in the field of metabolomics) where analytical techniques for the characterization of complex mixture samples have been developed. Through derivatization of functional groups with a labile proton using (e.g., trimethylsilyl), classes of compounds previously unsuitable for GC analysis become volatile and can be detected using fully automated sample preparation and analysis workflows (Khakimov et al., 2013). These analytical techniques open up for a 2-stage data analysis approach in the same analytical run: exploratory untargeted (unbiased) data analysis followed by a targeted analysis of a selection of compounds using a large sample set. These very low concentration compounds can be determined via a combination of automated sample pretreatment, GC, sensitive MS, and advanced data analysis. Regarding element analysis, inductively coupled plasma with optical emission spectroscopy (**ICP-OES**) has become more popular in recent years, providing a high-throughput, broad (unbiased) coverage both in terms of elements and concentration ranges (Hansen et al., 2009; Husted et al., 2011).

In the present study we characterize the chemical composition of UF, RO, ROP permeate, and storage tank water in a dairy ingredient production through untargeted derivatization based GC-MS coupled with advanced chemometric analysis. A selection of the chemical compounds are quantified via calibration standard series. Furthermore, detailed element composition of the streams is determined over a 10-h production run through ICP-OES analysis to investigate the dynamic behavior of element retention. This information was supplemented with 6-d growth experiments (radically surpassing normal process water storage) performed with 2 microorganisms isolated from RO permeate, a *Pseudomonas* sp. and a *Staphylococcus* sp., to test if the low nutrient levels in the ROP permeate were sufficient to support microbial growth. To the best of our knowledge this is the first study performed for an untargeted chemical characterization and element analysis of dairy membrane permeates.

MATERIALS AND METHODS

Samples Collected

All samples were collected at the Arla Foods Ingredients production facilities (Nr. Vium, Denmark) in 250-mL amber, sterile, polypropylene sample bottles (Isolab, Wertheim, Germany) over a sample collection period of less than 1 min, and stored at 5°C until analysis. Immediately before sampling, the process valves were opened and flushed to waste for approximately 10 s. Conductivity was measured on site, directly after collection, and before analysis for quality assurance purposes. A schematic illustration of the process and sampling locations is presented in Figure 1. Samples are denoted as follows: P1, UF permeate; P2, RO permeate; P3, ROP permeate before UV treatment (400 J/m^2, BX100e, Wedeco, USA); P4, ROP permeate after UV; P6, ROP permeate after storage tank, and UV treatment.

Process water samples for GC-MS analysis were collected 3 or 4 times on the same day and analyzed in duplicate or triplicate for P1, P2, P3, P4, and P6, leading to 15 samples overall (and 46 analysis runs in total). Process water samples for ICP-OES were collected simultaneously from UF, RO, and ROP permeate (P1, P2, and P4, respectively; 40 × 3 permeate samples; for further details, see Skou et al., 2017a). Sampling was initiated immediately after the treatment plant was started (following a cleaning in place) and continued for approximately 10 h. This time period included 2 feed tank changes where samples were collected with a higher frequency, accompanied by a lower sampling frequency in between.

The 2-stage RO+ROP plant consists of the RO section with 9 loops in parallel each with 6 membrane elements in series and the ROP segment of 4 loops in parallel each with 6 membrane elements in series; all elements are DOW Hypershell RO-8038 (TetraPak, Silkeborg, Denmark). The feed to the RO+ROP plant is UF permeate from whey processing adjusted to pH 5.8 with a flow ranging from 100 to 150 m^3·h^{-1}, target

retentate concentration is 15° brix (~13.5% DM). It should be noted that samples taken from the different extraction points at the same time are representative for averages and trends, but a sample-by-sample comparison between locations is not feasible due to different volumes and residence times at the various stages in the treatment plant.

Gas Chromatography-Mass Spectrometry

For P1 samples, a 100-μL aliquot was dried in a 200 μL glass insert using a ScanVac (LaboGene, Lynge, Denmark). For all other sample points a 1-mL aliquot was dried in a new glass vial using the ScanVac. The residues in the glass vials were re-suspended in 120 μL of Milli-Q water and a 100-μL aliquot out of this was further dried in 200-μL glass inserts. After drying, the inserts were capped in GC vials before derivatization and GC-MS injection. Dilution series of authentic standards were prepared in a concentration range of 64.9 to 0.4125 μM for lactose, glucose, and galactose, and 32.5 to 0.2125 μM for oleic acid (C18:1) and linoleic acid (C18:2). Thirty microliters of each standard solution was completely dried in the ScanVac and prepared for the GC-MS analyses as described previously. The different and distinct volumes for sample and standard preparation steps (dilution factors) are corrected for in the reported results.

Dried samples were derivatized by the addition of 30 μL of derivatization reagent trimethylsilyl cyanide (TMSCN, Fluka, Steinheim am Albuch, Germany) and agitation at 750 rpm for 50 min as described earlier in literature (Khakimov et al., 2013). The GC-MS analysis was performed using an Agilent 7890B GC (Agilent Technologies, Santa Clara, CA) coupled with a HT Pegasus time-of-flight mass spectrometer (Leco Corporation, USA). The GC-MS data acquisition parameters were published previously in Khakimov et al. (2013). Some modifications were introduced in this study including the GC oven temperature gradient and MS settings: the initial temperature of the oven was 40°C, held for 2 min, heated to 220°C at a rate of 20°C·min^{-1}, followed by 10°C·min^{-1} to reach a final temperature of 320°C at which the oven was kept for 8 min. The mass spectra were recorded in the range 45 to 500 m/z at a data acquisition rate of 10 spectra·s^{-1}. The raw GC-MS data were exported in the netCDF file format. The GC-MS data were processed using the freeware program PARADISe developed in our research group (www.models.life.ku.dk\PARADISe, accessed February 2017). This software enables deconvolution of peaks by means of PARAFAC2 (Johnsen et al., 2017). The PARAFAC2 (PF2) deconvoluted mass spectra were compared against the NIST11 GC-MS database version 2.0 [National Institute of Standards and Technology (**NIST**), Gaithersburg, MD]. The PF2 scores of deconvoluted peaks were extracted and used for subsequent data analysis. Scores from PF2 represent the normalized area of deconvoluted mass spectra, or in other words, the scores represent the normalized area of the specific compound extracted (Johnsen et al., 2017). All compounds were normalized with the score of a derivatization-agent-derived stable peak to correct for small differences in injection volume. Limit of detection (**LOD**) was set as a score value above the mean of the blanks plus 5 times the standard deviation of the blanks. If a specific compound was not included in a standard, that standard also served as blank for this compound, jointly with the 3 analytical blanks.

Inductively Coupled Plasma–Optical Emission Spectroscopy

Before analysis samples were acidified to 3.5% HNO_3 using 70% HNO_3 acid (Plasma-Pure, SCP Science, Marktoberdorf, Germany) to ensure fast washout in-between samples. Multi-elemental analysis was hereafter performed on a 5100 ICP-OES (Agilent Technologies) equipped with a Meinhard nebulizer and a cyclonic spray chamber. For each sample 22 elements were measured simultaneously. A 10-point external calibration standard from CPI International (Amsterdam, the Netherlands) was also included. The P1 samples were measured undiluted, and 10× and 100× diluted, whereas all other samples were measured undiluted. It is known that carbon in the samples will affect the plasma and change the plasma temperature. This changes how and when elements are ionized and eventually excited in the plasma (Husted et al., 2011). For this reason samples are typically digested before analysis to decrease the carbon to a negligible amount. However, this is a time-consuming process that can be circumvented if matrix matching is possible or if it can be shown that the matrix does not affect the analysis. To evaluate the actual interference from carbon, in our case almost exclusively lactose, the 10-point calibration curve was mixed with increasing levels of lactose (0, 0.04, 0.40, and 4.00% wt/wt). After each 20 (randomized) samples, a set of analyses consisting of a blank, a drift sample, and a blank were run to ensure that no drift or carry-over would affect the subsequent measurements (Olsen et al., 2016). The drift sample used was a certified biological reference material (NIST 1515, USA), certified for the elements presented in the Results section.

From the available ICP-OES spectral lines, one was chosen for each element, using sensitivity to lactose and drift over the analysis run as selection criteria. Calibrations were built based on the most appropriate standards (i.e., those standards that covered the sample's expected concentration range as best as possible). While predicting the concentrations in P1, the dilution with the signal best covered by the calibrations was used. The limits of quantification (**LOQ**) were established on 37 blank samples, calculated as the mean of the blanks plus 10 times the standard deviation of the blanks for each element. All observations below the quantification limits were excluded from further data evaluation.

To quantify the migration behavior of the different elements in the membrane systems the rejection factor (RF) was used:

$$\mathrm{RF} = \left(1 - \frac{\mathrm{C_{permeate}}}{\mathrm{C_{feed}}}\right) \cdot 100\%,$$

where $\mathrm{C_{feed}}$ and $\mathrm{C_{permeate}}$ denote the concentration for the feed and permeate, respectively.

Microbial Growth Potential

Microorganisms were isolated from UF and RO permeate by spreading permeates on plate count agar and water plate count agar and incubating at 22 and 37°C, respectively. Selected isolates were grown in urea broth (0.1 g·L^{-1} of yeast extract, 9.1 g·L^{-1} of KH_2PO_4, 9.5 g·L^{-1} of Na_2HPO_4, 20 g·L^{-1} of urea, and 0.01 g·L^{-1} of phenol red, Merck, Kenilworth, NJ) at 30°C to test their ability to degrade urea (data not shown). Two urease-positive isolates were chosen to test their growth potential in ROP permeate. The isolates were identified by 16S rDNA sequence analysis to be *Pseudomonas* sp. and *Staphylococcus* sp., respectively. The growth potential in ROP permeate was studied by inoculating cells (10^5 cfu·mL^{-1}) washed with sterile physiological saline in sterile filtrated ROP permeate (P3 in Figure 1) and calculating cfu·mL^{-1} during storage at 16°C for up to 6 d. The temperature 16°C was chosen because it is close to the temperature of the permeate during processing and the process water during storage, and thus the most relevant when investigating safety.

Statistical Analysis

Data were analyzed in Matlab R2015b (The MathWorks Inc., Natick, MA) using in-house routines.

RESULTS

Gas Chromatography-Mass Spectrometry

Compounds detected in the permeate samples are presented in Table 1, and some raw chromatograms of representative samples are shown in Figure 2. Table 1 illustrates at which sample point(s) the compounds were detected above the LOD. Detection (+) was marked if at least one of the analyses from the 3 distinct sample replicates' collected for each sample point contained the compound above the LOD. Compound identification was categorized into 3 groups: level 1 identification was based on spectral similarity and retention time comparison with pure standards, level 2 was based on spectral similarity >800 and a Kovats index (van Den Dool and Kratz, 1963) within less than 30 units from the reported values in the NIST library, and level 3 was based only on spectral similarity >800 (Sumner et al., 2007).

The semi-quantification of the different samples based on peak area represented by PARAFAC2 scores for α-ketogluterate ($C_5H_6O_5$, also known as α-ketogluteric acid in the protonated form) is shown as an example in Figure 3. The standards and blanks show a relative concentration consistently close to score zero and P1 shows the highest values, accompanied by a relative large sample-to-sample variance. The most abundant peak by far in the GC-MS profiles for P2, P3, P4, and P6 samples corresponded to urea. This chemical compound has been quantified in this production process in a previous study (Skou et al., 2017a) and was omitted from the GC-MS interpretation due to severe peak overloading making it unsuitable for quantification.

All calibration curves developed for the selected compounds had squared Pearson correlation coefficients of $r^2 = 0.88$ or higher. The lowest concentration of standards used for calibration curves are shown in Table 1. It was possible to quantify glucose ($r^2 = 0.95$) in P1, and lactose ($r^2 = 0.94$) and galactose ($r^2 = 0.95$) in P2. Glucose concentration in P1 samples was estimated to be between 0.01 and 0.06 m*M*, whereas lactose and galactose concentrations were estimated to be 0.02–0.10 m*M* and 0.005–0.02 m*M*, respectively, in P2 samples (Figure 4). Lactose and galactose concentrations in P1 samples were above the highest concentration standard (0.0649 m*M*) used in this study, whereas all other samples/sampling locations were below the LOD. The concentration of linoleic acid in the process water samples was above LOD and close to the lowest concentration standard. Oleic acid was not found above the LOD in any process samples.

Inductively Coupled Plasma-Optical Emission Spectroscopy

Calibration models for ICP-OES were tested for lactose interference (relevant especially for P1 samples), drift and linearity in the response range relevant for the samples. The final ICP-OES quantifications are presented in Table 2 as averages for each sampling point together with the number of measurements below the quantification limit, the average rejection factor over the RO- and ROP-membranes and the LOQ associated with the calibration for each observed element. As a point of reference the element composition of NIST Reference Material 1640a–Trace Elements in Natural Water is also included in Table 2.

Figure 5a presents the concentration of zinc as an example of the dynamic behavior and variance during the observed production run. The rejection for zinc per time point is presented in Figure 5b, and the averages of the 2 time series are included in Table 2.

To emphasize the connectivity between the elements found in process water samples, the correlation matrix, including measured conductivity, can be found in Table 3; the covariation between magnesium and conductivity as an example case is shown in Figure 6. Conductivity was on average reduced by 91.2% going from P1 to P2 and 82.2% going from P2 to P4.

Microbial Growth Potential

Figure 7 shows the microbial growth potential, indicating that *Pseudomonas* sp. was able to grow in ROP permeate, whereas *Staphylococcus* sp. was not. A log increase within 2 d was observed for *Pseudomonas* sp.

DISCUSSION

Gas Chromatography-Mass Spectrometry

Our GC-MS data analysis approach included (I) an exploratory untargeted approach where several dairy associated compounds were found (Table 1), and (II) a targeted quantification of selected compounds (Figure 4). With this approach, potential target molecules could be identified in P1, thereby making analysis of process water further along the process more focused

Table 1. Chemical compounds detected in process water samples arranged according to functional group[1]

No.	Compound[2]	KI NIST[3]	KI exp.[3]	P1	P2	P3	P4	P6	Lowest concentration standard (μM)
Organic acids									
10	Citric acid (2)	1,839	1,814	+	−	−	−	−	NA[4]
4	L-Malic acid (2)	1,538	1,516	+	−	−	−	−	NA
Sugars									
15	Lactose (1)	2,611	2,615	+	+	−	−	−	0.4125
11	Galactose (1)	1,846	1,846	+	+	−	−	−	0.4125
12	Glucose (1)	1,934	1,948	+	−	−	−	−	0.4125
13	Hexose (3)	∼1,900	1,979	+	−	−	−	−	NA
8	Pentose (3)	∼1,750	1,728	+	−	−	−	−	NA
Fatty acids									
14	Linoleic acid (1)	2,179	2,201	+	−	−	−	−	0.2125
—[5]	Oleic acid (1)	2,183	—	−	−	−	−	−	0.2125
3	Capric acid (2)	1,455	1,485	−	−	+	+	+	NA
7	Lauric acid (2)	1,651	1,673	+	+	+	+	+	NA
2	Glycerol (2)	1,300	1,288	−	+	−	−	−	NA
9	Glycerol-P (2)	1,744	1,759	+	+	−	−	+	NA
AA									
6	L-Glutamic acid (2)	1,629	1,642	+	+	+	+	+	NA
5	α-Ketoglutarate (2)	1,580	1,598	+	+	+	+	+	NA
Miscellaneous									
1	Phosphate (2)	1,285	1,296	+	+	+	+	+	NA
—[5]	Urea (1)	—	—	+	+	+	+	+	NA

[1]Identification numbers correspond to peak numbers in Figure 2. − = signals found to be less than the limit of detection (LOD; calculated as the mean of the blanks plus 5 times the SD of the blanks) for all samples; + = signals found larger than LOD in at least one measurement from a given sampling point.

[2]Numbers in parentheses indicate 1 = level 1 identification; 2 = level 2 identification; 3 = tentative identification.

[3]NIST = National Institute of Standards and Technology. Kovats index (KI) based on direct injection of 1 μL of alkane mixture (C10 to C40 from Sigma Aldrich, St. Louis, MO).

[4]NA = not available.

[5]Compound not detected in process water samples.

because compounds present in P1 are more likely to be present (in reduced concentrations) downstream.

In the initial exploratory analysis, compounds not included for targeted analysis were found (Table 1). Compounds that were not quantified using calibration curves are reported as either not detected (hence, below LOD) or detected (above LOD) only. Citric and malic acid are related to the primary energy metabolism of lactic acid bacteria in, for example, cheese making (Von Wright and Axelsson, 2012), whereas glycerol and glycerol-phosphate can be derived from either the primary energy metabolism or from fatty acid catabolism (Von Wright and Axelsson, 2012). L-Glutamic acid and its corresponding α-keto-acid, α-ketoglutarate, are important amino donors and receivers, respectively, in AA catabolism in cheese making and maturation (Ardö, 2006). The phosphate is likely from the phosphoric acid used to adjust pH before membrane filtration. Milk contains high amounts of the fatty acid triglycerides and free fatty acids are therefore expected to be present in the whey in low concentrations. Lactose and its breakdown products, glucose and galactose, constitute the primary content of the UF permeate and are thus anticipated at minor concentrations in downstream RO(P) permeates. To retain high signals for the low concentration compounds present in perme-

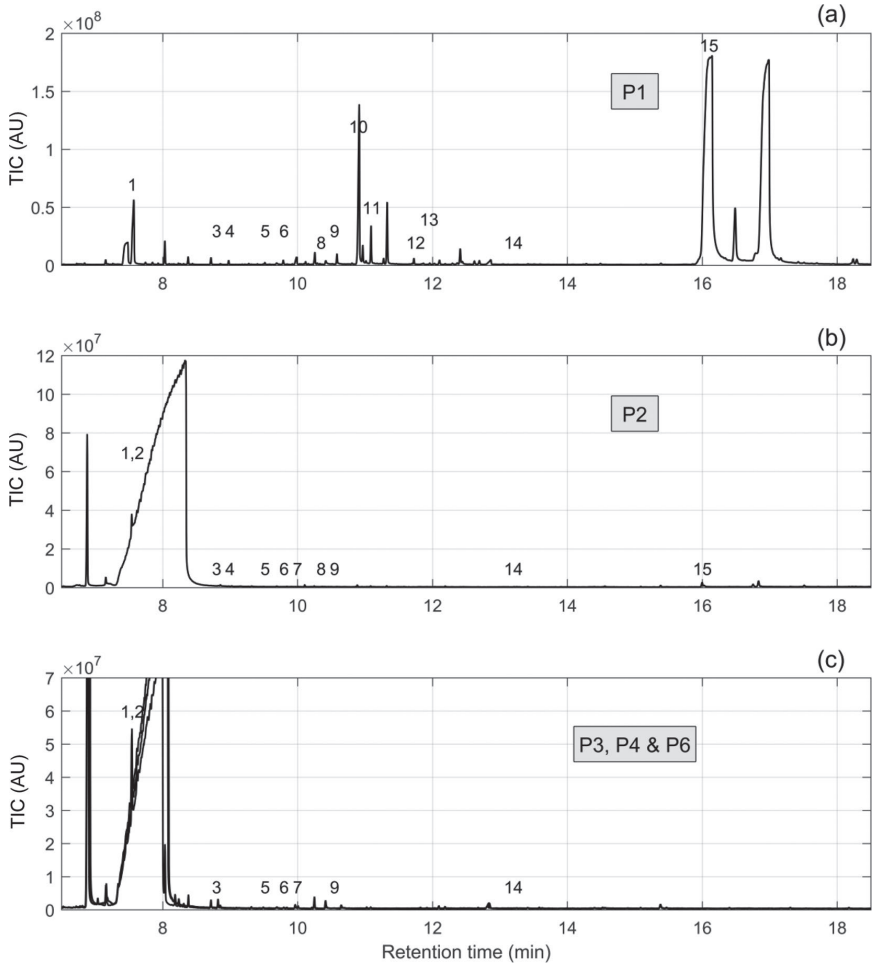

Figure 2. Examples of total ion count (TIC) chromatograms for single, representative (a) P1, (b) P2, and (c) P3, P4 plus P6 samples, with peak numbers corwresponding to the compounds in Table 1. The process water sampling points are marked P1 to P6. AU = arbitrary units.

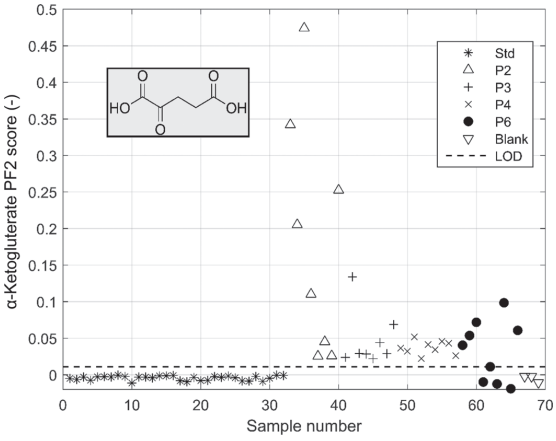

Figure 3. Relative concentration [PARAFAC2 (Johnsen et al., 2017; PF2) scores] of α-ketogluterate in all samples and standards. Relative concentrations of process water sampling point 1 (P1) samples (not shown) were between 1 and 9.2. LOD = limit of detection. Std = standard.

ates, samples were trimethylsilylated directly, without prior methoximation. As a result, the unexpected peaks of hexose and pentose may derive from glucose and galactose (Koek et al., 2006). Without methoximation, several peaks corresponding to stereoisomers of carbohydrates (Eversheld, 1993) make unambiguous identification almost impossible when complex samples are investigated. Urea has been known to pass RO membranes and the concentrations in the permeate using the present membrane processing system were reported earlier (Skou et al., 2017a). Glucose was quantified in P1 samples (0.01–0.06 mM), whereas lactose and galactose were quantified in P2 samples (0.02–0.10 mM and 0.005–0.02 mM, respectively).

When looking at Table 1, it is important to bear in mind that the samples collected simultaneously at points P1 to P4 (Figure 1) are not one-to-one comparable due to the large flows involved in parallel membrane batteries and the unknown hold-up times, whereas P6 samples are drawn from a storage tank. There will be an element of chance and intrinsic variability over time when sampling large industrial-scale systems, and one can only postulate that the samples collected are representative of the (average) process performance. Looking at the RO(P) permeates (P3, P4, and P6), it was found that fatty acids, glycerol, AA, and phosphate pass the membrane to some extent. Interestingly, large molecules such as fatty acids seem to be able to pass the RO(P)-membranes, albeit in low concentrations. This observation supports a recent study by Cortés-Francisco and Caixach (2013) who investigated a seawater desalination process and also found fatty acids permeating the RO-membranes. It seems feasible that glycerol, the smallest AA glutamic acid and the corresponding α-keto acid and phosphate, would be able to permeate the membranes in minute concentrations; however, to our knowledge no studies have reported this before.

Glycerol is detected in the RO(P) permeates, but not in the UF permeate, probably due to the higher concentration factor of the RO(P) samples. This could indicate that the glycerol concentration is more or less constant in the process water streams, suggesting that it can pass RO-membranes, as was also observed for urea (Skou et al., 2017a). However, this should be investigated further to get more solid confirmation. Phosphate was found in all process water streams.

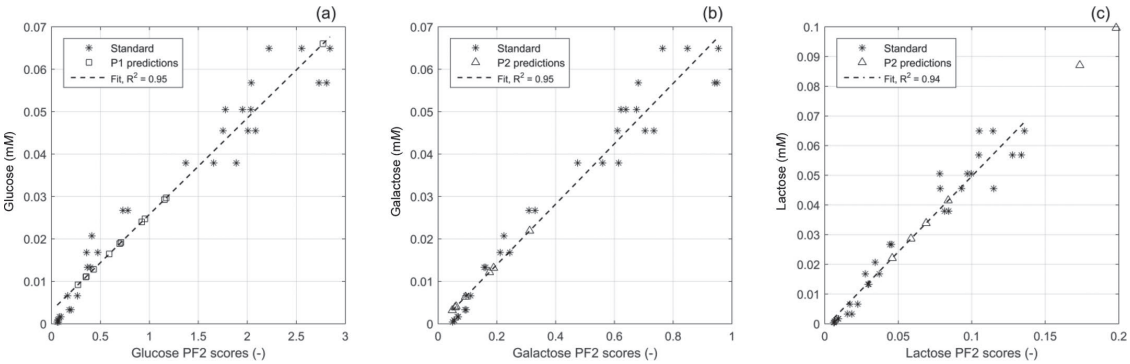

Figure 4. Quantification based on relative concentrations [PARAFAC2 (Johnsen et al., 2017; PF2) scores] and standard series for (a) glucose in process water sampling point 1 (P1) samples, and (b) galactose and (c) lactose in process water sampling point 2 (P2) samples.

Table 2. Overview of elements detected in process water samples; the average concentration (based on samples with levels above the limit of quantification) in each process water sampling point; rejection factors over the reverse osmosis and reverse osmosis polisher membranes

Element (LOQ[1])	NIST ref.[2]	Concentration ($\mu g \cdot L^{-1}$) [measurements below LOQ]						Rejection factor	
		P1 [n = 40]		P2 [n = 39]		P4 [n = 40]		P1 to P2	P2 to P4
Al (2.7 $\mu g \cdot L^{-1}$)	53.0	17.2	[0]	3.1	[32]	3.0	[38]	82.3	—
B (5.6 $\mu g \cdot L^{-1}$)	303.1	92.7	[0]	65.0	[0]	38.1	[0]	29.2	41.3
Ca (25.0 $\mu g \cdot L^{-1}$)	$5.6 \cdot 10^3$	$206 \cdot 10^3$	[2]	395	[1]	—	[40]	99.8	—
Cr (0.58 $\mu g \cdot L^{-1}$)	40.5	6.34	[37]	—	[39]	—	[40]	—	—
Cu (1.4 $\mu g \cdot L^{-1}$)	85.8	2.1	[38]	1.8	[35]	1.9	[39]	—	—
Fe (5.3 $\mu g \cdot L^{-1}$)	36.8	6.9	[16]	—	[39]	—	[40]	—	—
K ($2.42 \cdot 10^3$ $\mu g \cdot L^{-1}$)	579.9	$799 \cdot 10^3$	[4]	$32 \cdot 10^3$	[1]	$5.6 \cdot 10^3$	[3]	96.0	82.7
Mg (-1.1 $\mu g \cdot L^{-1}$)[3]	$1.06 \cdot 10^3$	$40.3 \cdot 10^3$	[0]	86.4	[0]	4.1	[0]	99.3	93.8
Mo (1.6 $\mu g \cdot L^{-1}$)	45.6	11.7	[3]	—	[39]	—	[40]	—	—
Na (200 $\mu g \cdot L^{-1}$)	$3.14 \cdot 10^3$	$115 \cdot 10^3$	[0]	$3.41 \cdot 10^3$	[0]	590	[0]	96.5	81.5
P (38.0 $\mu g \cdot L^{-1}$)	—	$232 \cdot 10^3$	[0]	614	[1]	—	[40]	99.4	—
S (110 $\mu g \cdot L^{-1}$)	—	$26 \cdot 10^3$	[0]	66	[1]	15.0	[8]	95.1	77.3
Zn (3.7 $\mu g \cdot L^{-1}$)	55.6	33.9	[0]	8.7	[1]	8.2	[0]	70.0	−0.6

[1]Limits of quantification (LOQ) were calculated as the mean of the blanks plus 10 times the SD of the blanks for each element.

[2]National Institute of Standards and Technology (NIST) Standard Reference Material 1640a.

[3]Negative LOQ is a result of the large span in the standards causing the model to show a small bias in the very low concentration range.

Comparing this with the ICP-OES results, where P4 samples are below the detection limit (38 $\mu g \cdot L^{-1}$), suggests that the concentration permeating is very low indeed. The compounds found in the ROP permeates were all in very low concentrations (except urea) and on the border of the detection limit. Nevertheless, the

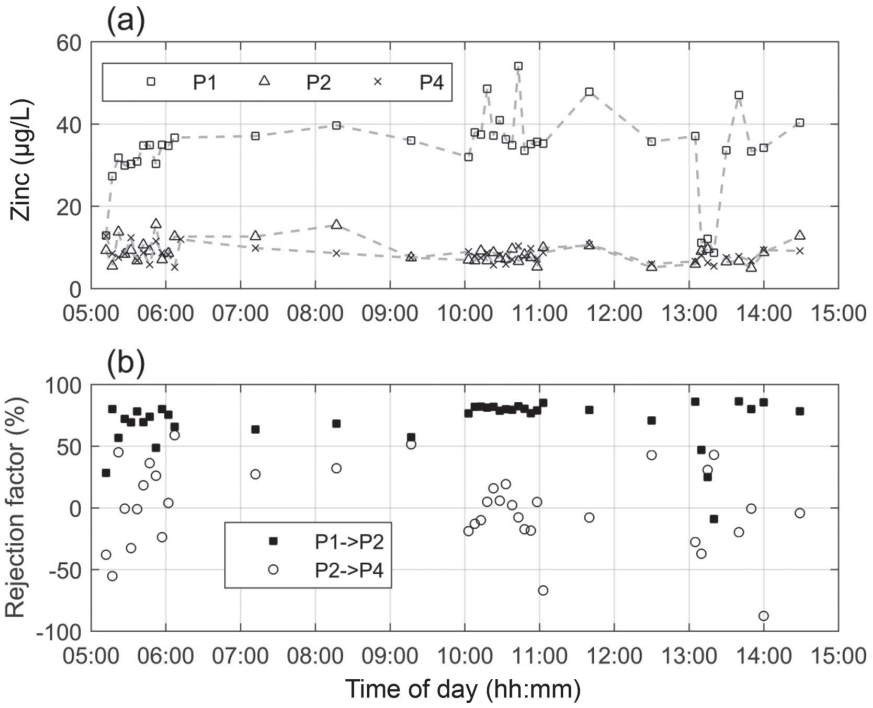

Figure 5. (a) Zinc concentration at the 3 sampling locations and (b) zinc rejection factor over a 10-h processing period.

analysis does suggest that some compounds beside urea do permeate the ROP-membranes.

We attempted to perform unbiased GC-MS analysis and employed limited sample pre-cleaning and extraction. The GC-MS approach applied in this study was focused to detect compounds with molecular mass of up to 1 kDa. The compounds must also be volatile and thermally stable to be detected by our method. The derivatization usually enhances the above-mentioned 2 characteristics, enabling better detection. However, the drying stage involved in this protocol may result in evaporation of naturally volatile compounds such as ethanol and short-chain fatty acids.

Inductively Coupled Plasma-Optical Emission Spectroscopy

As described earlier, it is known that carbon in the samples will affect the plasma and change the plasma temperature. In this study we chose not to digest samples as this is a time-consuming process; instead, lactose matrix interference was tested. None of the spectral lines selected were found to be sensitive to lactose (carbon) at low levels (0.04% wt/wt).

From the total 22 elements available in the ICP-OES procedure, 13 elements were sensitive enough to quantify in the process water samples as presented in Table 2. Chrome, copper, iron, and molybdenum levels were so minimal that only very few samples were above the limit of quantification; this also means that rejection calculations could not be performed. For aluminum 7, P2 samples were above the detection limit and the rejection calculation was performed, which should, however, be interpreted with caution based on the small number of observations. Only 2 P4 samples were above the LOQ and the result is reported here only to illustrate the levels found in some process water samples. Our findings support the results from Vourch et al. (2008) for Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, and HPO_4^{+} rejection in RO permeate, where we found >99, >99, 94–99, 87–98, and >99%, respectively. The negative detection limit for magnesium is an artifact of the calibration, which is focused on the expected range of the samples, leading to a poor estimation of the blank values. For zinc a 70% rejection is obtained from the first RO-membrane, but no additional reduction appears over the ROP-membrane. From Figure 5, the dynamics of the process becomes clear, especially around the tank change at 10:00 h where more extreme observations are made. This type of information could lead plant managers and optimization engineers to investigate extreme behavior of the process (Skou et al., 2017b). Sulfur is rejected effectively over both membrane plants, leading to a rejection efficiency of 95 and 77%, respectively. Boron was detected in all samples (also as a consequence of the ICP-OES method being very sensitive toward this element) and results show a modest rejection efficiency of the RO(P)-membranes (29.2 and 41.3%, respectively). Similar rejection efficiencies have been reported by Rodríguez Pastor et al. (2001), who also hypothesize that the low efficiency is due to boron being in the form of boric acid, which has no ionic charge, making it permeate the membranes much like water. Despite the poor rejection, the concentrations found at 92.7, 65.0, and 38.1 $\mu g \cdot L^{-1}$ (with an average SE across all predictions of 0.44 $\mu g \cdot L^{-1}$) for UF, RO, and ROP permeate, respectively, are far below the maximum guideline values of $2.4 \cdot 10^3$ $\mu g \cdot L^{-1}$ specified by the World Health Organization (WHO, 2009) for potable water, a conclusion that holds for all elements in the P3 stream. Aluminum was in a low concentration in P1 (17.2 $\mu g \cdot L^{-1}$) and was reduced close to the detection limit (2.7 $\mu g \cdot L^{-1}$) for P2.

From the correlation matrix presented in Table 3, it can be seen that all the elements are highly correlated

Table 3. Squared Pearson correlation (r^2) between all elements and conductivity (Cond.) over all measurements; Cr, Cu, and Fe are not reported (—) due to less than 20% of the observations being available for calculation, as decided from the limits of quantification (calculated as the mean of the blanks plus 10 times the SD of the blanks for each element)

Item	Al	B	Ca	Cr	Cu	Fe	K	Mg	Mo	Na	P	S	Zn	Cond.
Al	1.00													
B	0.43	1.00												
Ca	0.92	0.71	1.00											
Cr	—	—	—	—										
Cu	—	—	—	—	—									
Fe	—	—	—	—	—	—								
K	0.91	0.72	0.99	—	—	—	1.00							
Mg	0.94	0.67	1.00	—	—	—	1.00	1.00						
Mo	0.31	0.26	0.50	—	—	—	0.23	0.33	1.00					
Na	0.90	0.68	0.99	—	—	—	1.00	0.99	0.12	1.00				
P	0.92	0.72	1.00	—	—	—	1.00	1.00	0.24	0.99	1.00			
S	0.92	0.71	0.99	—	—	—	1.00	1.00	0.13	0.99	1.00	1.00		
Zn	0.82	0.64	0.91	—	—	—	0.90	0.92	0.51	0.90	0.89	0.90	1.00	
Cond.	0.95	0.70	0.99	—	—	—	1.00	1.00	0.26	0.99	1.00	1.00	0.91	1.00

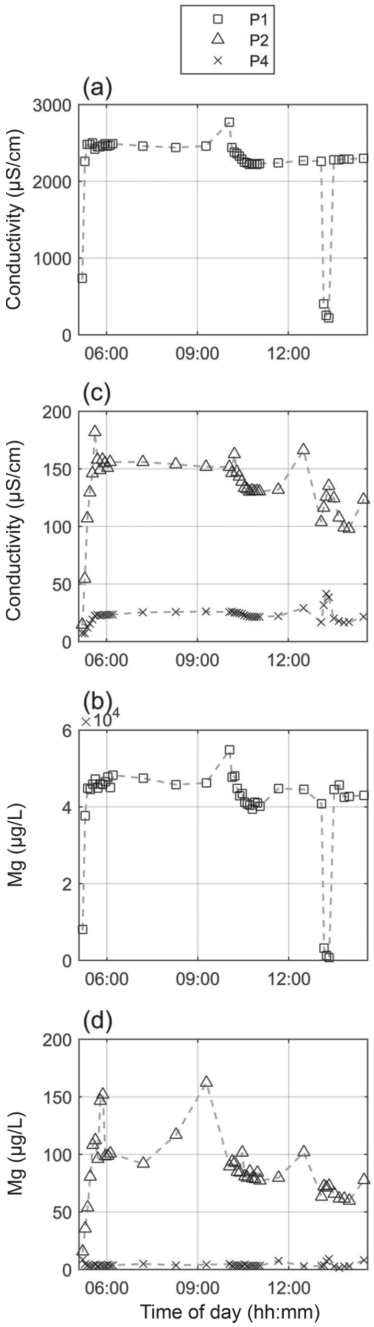

Figure 6. (a, c) Conductivity values and (b, d) magnesium concentration for all samples from sampling points (a, b) P1 and (c, d) P2 plus P4. The process water sampling points are marked P1 to P4.

except for boron and molybdenum, where it should be reminded that the former is not rejected effectively and the latter was quantified only in a few measurements and these were close to the LOQ. Conductivity correlates strongly with all other elements. This suggests that conductivity can be safely used as an indirect measure of the element concentration, as is done in daily dairy operational practice. The constant rejection profiles over time in Figure 5 indicate that the process was running very stable, except during, for example, tank change, for the sampled period and that our findings are thus representative of normal operating conditions (Skou et al., 2017a).

Microbial Growth Potential

Although ROP permeate contains very low levels of bacteria, there is still a risk that undesirable growth may occur if some of the population is able to proliferate. Out of 2 urease positive strains, a *Pseudomonas* sp. and a *Staphylococcus* sp. isolated from RO permeate, only the *Pseudomonas* sp. was able to grow. The inoculation levels were several logs higher than found in any permeate and growth seemed to cease around 5×10^6 cfu·mL^{-1}. The observations nevertheless indicate that even the very low nutrient ROP permeate may support growth at 16°C of part of the microbial population present. The *Staphylococcus* sp. survived but was not able to grow in the ROP permeate despite its urease activity, suggesting that additional nutrients were needed for the growth of this organism.

CONCLUSIONS

This investigation presents novel untargeted and targeted methodologies to characterize the chemical composition of process water derived from membrane processes in a dairy ingredient production facility. This first unbiased look into the chemical composition forms the input for a measurement-based safe use of reclaimed process water in the dairy industry. The methodologies presented facilitate further investigations potentially leading to periodic, high-sample-number risk assessment of process water and use potential to guarantee a hygienic and safe process operation. Results suggest that bigger compounds (long-chain fatty acids, glycerol-phosphate, and glutamic acid) do permeate the ROP-membranes but in very minute concentrations, whereas the elements boron and urea permeates readily. Although the nutrient levels were very low, *Pseudomonas* sp. originating from process water was able to grow in ROP permeate stressing the importance of controlling temperature and maximum-allowed storage times of the permeate intended for use.

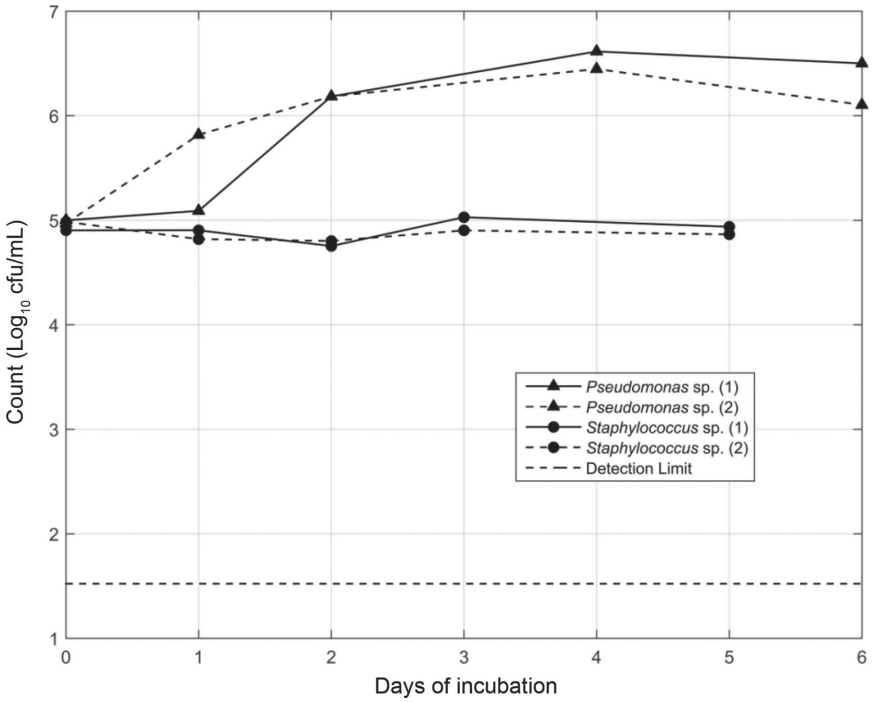

Figure 7. Growth curves of *Pseudomonas* sp. and *Staphylococcus* sp. in reverse osmosis polisher permeate at 16°C.

ACKNOWLEDGMENTS

This work was supported by the REuse of WAteR in the food and bioprocessing inDustries (REWARD) consortium funded by the Danish Council for Strategic Research, Programme Commission on Health Food and Welfare (Copenhagen).

REFERENCES

Ardö, Y. 2006. Flavour formation by amino acid catabolism. Biotechnol. Adv. 24:238–242. https://doi.org/10.1016/j.biotechadv.2005.11.005.

Casani, S., M. Rouhany, and S. Knøchel. 2005. A discussion paper on challenges and limitations to water reuse and hygiene in the food industry. Water Res. 39:1134–1146. https://doi.org/10.1016/j.watres.2004.12.015.

Cortés-Francisco, N., and J. Caixach. 2013. Molecular characterization of dissolved organic matter through a desalination process by high resolution mass spectrometry. Environ. Sci. Technol. 47:9619–9627. https://doi.org/10.1021/es4000388.

Evershed, R. P. 1993. Advances in Silylation. K. Blau and J. M. Halket, ed. John Wiley & Sons, Hoboken, NJ.

Hansen, T. H., K. H. Laursen, D. P. Persson, P. Pedas, S. Husted, and J. K. Schjoerring. 2009. Micro-scaled high-throughput digestion of plant tissue samples for multi-elemental analysis. Plant Methods 5: https://doi.org/10.1186/1746-4811-5-12.

Husted, S., D. P. Persson, K. H. Laursen, T. H. Hansen, P. Pedas, M. Schiller, J. N. Hegelund, and J. K. Schjoerring. 2011. Review: The role of atomic spectrometry in plant science. J. Anal. At. Spectrom. 26:52–79. https://doi.org/10.1039/C0JA00058B.

Johnsen, L. G., P. B. Skou, B. Khakimov, and R. Bro. 2017. Gas chromatography–mass spectrometry data processing made easy. J. Chromatogr. A 1503:57–64. https://doi.org/10.1016/j.chroma.2017.04.052.

Khakimov, B., M. S. Motawia, S. Bak, and S. B. Engelsen. 2013. The use of trimethylsilyl cyanide derivatization for robust and broad-spectrum high-throughput gas chromatography-mass spectrometry based metabolomics. Anal. Bioanal. Chem. 405:9193–9205. https://doi.org/10.1007/s00216-013-7341-z.

Koek, M., B. Muilwijk, M. J. van der Werf, and T. Hankemeier. 2006. Microbial metabolomics with gas chromatography/mass spectrometry. Anal. Chem. 78:1272–1281. https://doi.org/10.1021/ac051683+.

Olsen, L. I., T. H. Hansen, C. Larue, J. T. Østerberg, R. D. Hoffmann, J. Liesche, U. Krämer, S. Surblé, S. Cadarsi, V. A. Samson, D. Grolimund, S. Husted, and M. Palmgren. 2016. Mother-plant-mediated pumping of zinc into the developing seed. Nat. Plants 2:16036. https://doi.org/10.1038/nplants.2016.36.

Rodríguez Pastor, M., A. Ferrándiz Ruiz, M. F. Chillón, and D. Prats Rico. 2001. Influence of pH in the elimination of boron by means of reverse osmosis. Desalination 140:145–152. https://doi.org/10.1016/S0011-9164(01)00364-2.

Skou, P. B., T. A. Berg, S. D. Aunsbjerg, D. Thaysen, M. A. Rasmussen, and F. van den Berg. 2017a. Monitoring process-water quality using near infrared spectroscopy and partial least squares regression with prediction uncertainty estimation. Appl. Spectrosc. 71:410–421. https://doi.org/10.1177/0003702816654165.

Skou, P. B., S. E. Holroyd, and F. van den Berg. 2017b. Tutorial–Applying extreme value theory to characterize food-processing systems. J. Chemom. e2896. https://doi.org/10.1002/cem.2896.

Sumner, L. W., A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, and M. R. Viant. 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3:211–221. https://doi.org/10.1007/s11306-007-0082-2.

van den Berg, F., C. B. Lyndgaard, K. M. Sørensen, and S. B. Engelsen. 2013. Process Analytical Technology in the food industry. Trends Food Sci. Technol. 31:27–35. https://doi.org/10.1016/j.tifs.2012.04.007.

van Den Dool, H., and P. D. Kratz. 1963. A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. J. Chromatogr. 11:463–471. https://doi.org/10.1016/S0021-9673(01)80947-X.

Von Wright, A., and L. Axelsson. 2012. Lactic Acid Bacteria: An Introduction. 4th ed. S. Lahtinen, A. C. Ouwehand, S. Salminen, and A. Von Wright, ed. CRC Press, Boca Raton, FL.

Vourch, M., B. Balannec, B. Chaufer, and G. Dorange. 2005. Nanofiltration and reverse osmosis of model process waters from the dairy industry to produce water for reuse. Desalination 172:245–256. https://doi.org/10.1016/j.desal.2004.07.038.

Vourch, M., B. Balannec, B. Chaufer, and G. Dorange. 2008. Treatment of dairy industry wastewater by reverse osmosis for water reuse. Desalination 219:190–202. https://doi.org/10.1016/j.desal.2007.05.013.

WHO. 2009. Boron in drinking-water. Background document for development of WHO Guidelines for Drinking-Water Quality. World Health Organization, Geneva, Switzerland.

2.1 Lactose interference in ICP-OES

In **Paper I** the issue of carbon interference in ICP-OES analysis was raised since excessive amounts of organic material can lead to problems with nebulisation and plasma stability and carbon interferences creating polyatomic ions.[48] This can be overcome by either digesting the samples (i.e. oxidizing all carbon to carbon dioxide gas) or matrix matching the standards for quantification.

In order to establish whether my situation was problematic enough to necessitate matrix matching lactose was added in increasing concentrations up to 4% to the calibration standards. Lactose was by far the most abundant compound in UF permeate and therefore it was chosen as the single carbon source. The effect of lactose addition was not shown in **Paper I**, but is shown in Figure 2.5 where four spectral bands for calcium are presented. To the ten mixture standards of elements increasing amounts of lactose were added and it can be seen that the 612.222 nm line (d) shows slight intensity suppression with 4% lactose. The line at 396.847 nm (c) is so sensitive that only the three lowest concentration standards are within the range of the instrument. In the end the spectral line 318.127 nm (a) was chosen for quantification due to the low interference to lactose and linear response range.

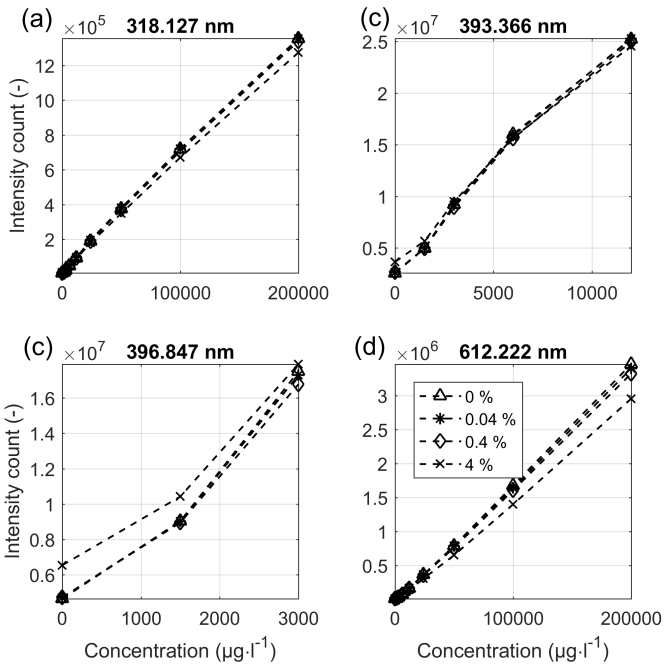

Figure 2.5: Intensities of four spectral bands (a–d) for calcium as function of concentrations calcium and increasing amounts of lactose.

2.2 Monitoring long-term performance

To establish the urea and lactose levels across several production runs samples were collected over a two week period, using a stratified randomization strategy (see Table 2.3). The aim was to estimate the natural compositional distribution / variability in the recovered water stream. Production schedules do not follow operator shifts, so the the idea was to collect three samples every day at random time-points. Due to the long and infrequent sampling plan the sample collection was left to the process operators on-site. Therefore, to relieve the work load, the timing on sample collection was relaxed to large time intervals (three hour windows) rather than exact times. This was judged sensible based on the large volumes involved, the block-wise nature of the operation (where standard stirred 250 m^3 tanks are used as both feed and receiver units) and the filtering or averaging-out effect this has.

Table 2.3: Initial sampling plan, 1 indicates a request for a sample in the given time interval.

Time/day	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
00-03	1	0	0	1	1	1	0	0	1	0	1	1
03-06	0	0	0	0	1	0	1	0	0	1	0	0
06-09	0	1	0	0	0	1	0	1	1	0	0	0
09-12	1	1	1	1	0	0	0	1	0	1	0	1
12-15	1	1	1	0	1	1	1	0	1	0	0	0
15-18	0	0	0	0	0	0	0	0	0	0	1	0
18-21	0	0	0	1	0	0	0	1	0	0	1	1
21-24	0	0	1	0	0	0	1	0	0	1	0	0

The lactose and urea concentration in the samples obtained were quantified using enzyme kits (Megazyme, IRE). The on-line brix measurements on the retentate were also retrieved from the production site. From this separate batches could be identified: high brix indicated production while low brix indicated no production (due to e.g. switch-overs, cleaning or maintenance). The lactose and urea results are shown in Figure 2.6. Comparing the samples obtained in Figure 2.6 with the requested samples in Table 2.3 illustrates the challenge of planning experiments on industrial plants. Of course the primary objective is to run the plant and extra activities thus often have to be abandoned. Urea was rejected inefficiently as seen in **Paper I** while lactose was mostly rejected efficiently (concentrations estimated below the detection limit of 1.5 ppm). On two occasions — October 30 06:00–09:00h and November 1 21:00–24:00h — lactose was quantified to 17 and 16 ppm, respectively. This indicates that sometimes lactose will penetrate the membranes. The reason for these two events is unknown.

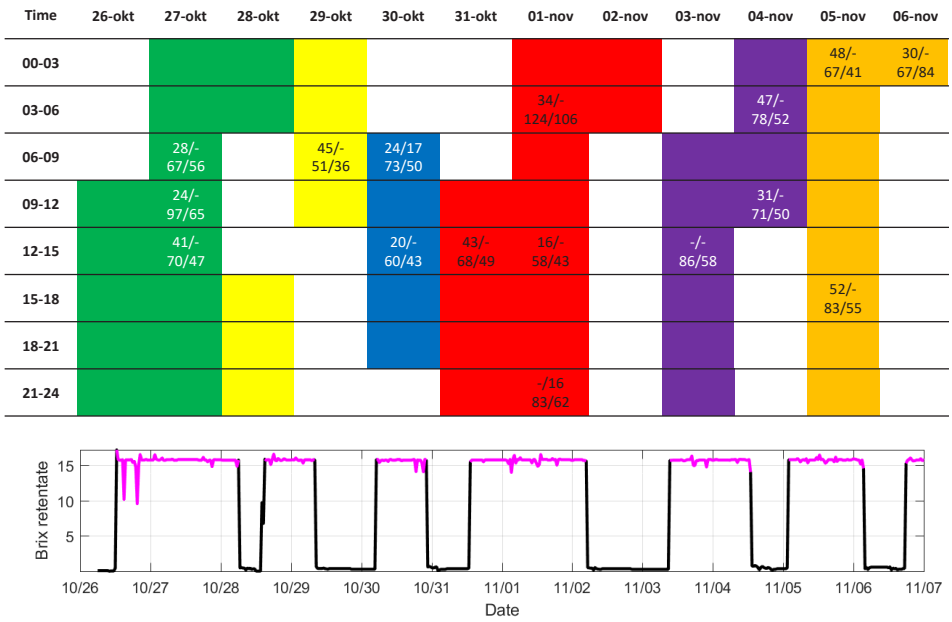

Figure 2.6: RO and ROP permeate samples collected using a stratified-random strategy during processing from October 26 till November 6, 2015. Distinct colors indicate individual batches / process runs identified from the on-line brix measurements on the retentate side (bottom graph), white indicates no processing / CIP. In each cell the top row contains RO / ROP values for lactose in ppm and the bottom row contains RO / ROP concentrations for urea in ppm; below detection limit indicated by a dash (-).

2.3 New developments in membrane filtration

In attempts to lower operating costs of RO filtration processes new technologies such as forward osmosis (FO) are emerging. In FO the principle is to draw the permeate through the membranes by adding a osmotic agent (e.g. a salt) to the permeate (the *draw solution*) rather than push the feed through with pressure.[49] This of course results in FO permeate with relative high concentration of the osmotic agent. This can be removed by a conventional RO process resulting in reclaimed water (RO permeate) and the retentate can be considered as regenerated osmotic agent and reused. Aydiner et al.[50] tested FO to reclaim water from whey and concluded that economically FO+RO was comparable to conventional UF+RO systems.

Another technological development to lower operating costs is to incorporate biological water channels, specifically aquaporin proteins, into the membrane material thereby creating biomimetic membranes.[51] Adding aquaporin proteins to membranes will ideally only transport water molecules over the membrane wall

thus lowering the pressure needed to obtain sufficient flux. The technology is still very new but the first reports are coming out. Kalafatakis et al.[52] investigated a water recovery strategy using FO and biomimetic membranes for a 2nd generation biorefinery. Crude glycerol and enzymatically treated wheat straw was used as the draw solution to concentrate the product stream. This resulted in a diluted draw solution, which was in fact the feedstock for the biorefinery, thereby creating a closed loop.

2.4 Processing of GC-MS data

In **Paper I** the PARAllel factor analysis2 based Deconvolution and Identification System (PARADISe) was used to process the raw GC-MS data. The software was developed after realising that most commercial and freely available programmes were inadequate in terms of inspecting data, looking up mass spectra or deconvoluting overlapping peaks or any combination hereof. A graphical user interface was built in Matlab® for inspecting and cutting the data in the retention time dimension (creating intervals) as well as inspecting mass spectra via the NIST MSSearch software. The output is a peak table of the deconvoluted areas for all samples and the tentative identifications from MSSearch. The deconvolution engine is, as the name suggests, PARAlell factor analysis2 (PARAFAC2). Box 4 on page 54 describes basic PARAFAC modelling and the main difference to PARAFAC2 is that the latter can handle retention time shifts (which are practically unavoidable in GC-MS), while still obtaining a unique result. The software is more thoroughly described in **Paper IV** and is freely available on `http://models.life.ku.dk/paradise`.

Chapter 3

Monitoring membrane permeate quality

In dairy processing research and development the main attention concerning membrane performance has been on separating the product (retentate) from the input streams. More recently, with attention being focussed towards sustainable processing, also the permeate quality has become of interest as shown in e.g. Table 2.2 in the previous chapter.

Membrane filtration processes are subject to disturbances during operation, just like any other process. Being able to react to these disturbances requires the ability to monitor the processing performance. One of the main disturbances in membrane filtration processing is compromised integrity (i.e. breaking of the membrane). There are two overall approaches to testing membrane integrity: direct or indirect testing.[53–55]

Direct testing is primarily pressure or vacuum testing of individual membrane elements. In the case of RO membranes vacuum testing is preferred since pressure testing requires pressure on the frail permeate side which may damage the membrane. The speed of the pressure / vacuum drop indicates whether a leak is present. When carrying out vacuum testing the ASTM standard prescribes draining the membranes for one hour and then soaking them in membrane permeate overnight before testing.[54] This means that direct testing is mostly used as a quality check before assembling the processing plants or when replacing old membrane units.

Indirect testing concerns measuring the integrity implicitly, i.e. on permeate quality. Techniques can be both on-line and off-line. Typical off-line techniques are microbiological challenge tests where micro organisms are introduced into the system and rejection efficiencies are measured.[54] However, off-line monitoring will typically not provide information in time for preventive actions, and it has been argued that they are more accurately classified as direct testing methods.[55] Rather the focus here will be on on-line methods.

On-line monitoring of RO membrane integrity has received modest attention in literature. Adham et al.[53] tested several on-line methods: conductivity, particle counting and TOC. They investigate whether the methods were sensitive enough to detect a difference in permeate quality between a normal running system and the same system with a purposely compromised O-ring.* TOC was found to be the most sensitive method, while conductivity was less sensitive but still able to detect the loss of integrity which on-line particle counting could not discover. They further speculate the need for more than one measurement technology to ensure reliable operation of an RO processing unit. Kumar et al.[54] reviewed methods for direct and indirect monitoring / testing of RO membrane integrity. The on-line methods included conductivity, TOC, particle counting and sulphate analysis. On-line sulphate monitoring is based on ion chromatography and is thus not nearly as fast nor cost-effective (due to high acquisition, training and maintenance cost) and is not further considered. On-line TOC analysers were reported to have improved sensitivity from 1 $\mu g \cdot l^{-1}$ to 0.03 $\mu g \cdot l^{-1}$.

More recently fluorescence spectroscopy has been proposed[56] and tested[57,58] as a potential monitoring tool for *recycled water systems* (which also covers RO membrane processing; for more information on fluorescence spectroscopy see Box 5 on page 61).

Singh et al.[57] used the naturally fluorescence of dissolved organic matter as the key quality attribute to ensure the membrane filtration process was running as intended. Despite the fluorescing dissolved organic matter being present in different water sources, no detailed characterisation of the permeate was performed. In a earlier review done by the same research group[56] it was found difficult to establish a link between fluorescence excitation-emission (EEM) landscapes and other bulk water quality parameters such as BOD, COD and TOC. As stated in the review, this is likely due to a varying proportionality of fluorescing to non-fluorescing compounds contributing to the TOC, meaning if the ratio between fluorescing / non-fluorescing compounds and TOC is not constant, they cannot be correlated directly. Their strategy thus assumes and relies on fluorescing dissolved organic matter being the key quality attribute.

Based on these references, no detailed chemical characterisation of the membranes have been done *a priori* to identify a specific key quality indicator, i.e. a representative chemical compound. Regarding on-line TOC and conductivity, these techniques measure *bulk* properties of the water in the sense that they hold no specific information about what compounds permeate. In the case of fluorescence the underlying indicator(s) of compromised integrity are identifiable up

*Gasket or washer ensuring membranes are tightly fixed inside the casing.

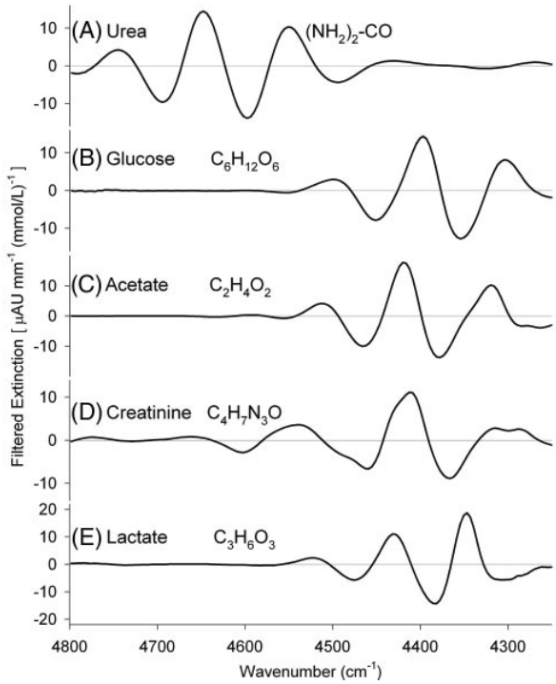

Figure 3.1: Pure component pre-processed near infrared spectra reproduced from Olesberg et al.[59]

to compound class, but the fact that the specific permeating compounds are unknown (fluorescent as well as non-fluorescent) lowers confidence that the method will detect an issue when it occurs.

While TOC and conductivity may be useful for measuring overall membrane integrity permeation of urea is not (or only slightly) related to membrane malfunctioning. Urea permeates RO membranes by another mechanism due to its close similarity to water. Since urea is the main organic compound permeating the membranes and it does not permeate like most compounds, it must be monitored purposefully.

Paper I documents that in the process-water from the dairy ingredient production, urea is clearly the main compound permeating the membranes. Quantification of urea has been a wide research topic within medical device development in order to determine when dialysis treatment is finished for a patient.[59–62] Interestingly, urea has a very specific signal in the near infrared (NIR) spectrum when compared to many other chemical compounds. However, there are still interferents, e.g. glucose in the case of dialysis and lactose in the case of ROP membrane permeate. From the dialysis literature very informative figures of urea's NIR spectrum (Box 3) were found shown in Figure 3.1.

In **Paper II** we seek to use NIRS to quantify urea as a key quality attribute in process-water recovered in the dairy ingredient industry.

Box 3: Near InfraRed Spectroscopy — NIRS

The NIR spectral region lies between the visual and infrared range. This region is characterised by overtone and combination bands of the fundamental vibrations from the infrared region. Fundamental molecular vibrations can be understood from a harmonic diatomic oscillator model. Since the vibration would be harmonic, the potential energy curve would in principle also be parabolic in shape and symmetrical as seen on the left hand side lower part in Figure 3.2 where the diatomic molecule is excited from ground state (n=0) to first excited state (n=1).

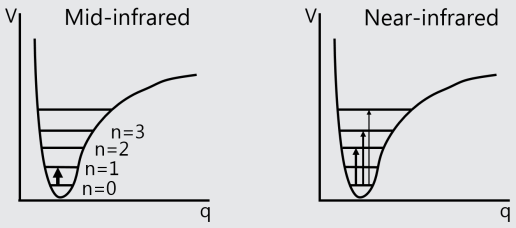

Figure 3.2: The harmonic and anharmonic oscillator can be used to describe infrared and near-infrared vibrations. V denotes the potential energy and q the displacement of the atoms. Illustration inspired by Siesler.[63]

The potential energy depends on the force constant of the bond and masses of the atoms. This means that bonds in molecules with different masses (as well as different bond types) will give rise to different fundamental vibrations. Vibrational energies can only take certain discrete values, i.e. certain overtones. In overtones the displacement increases compared to the fundamental frequencies with the result that the oscillations are no longer harmonic, but anharmonic and therefore not symmetric as shown on the right hand side of Figure 3.2. This is due to repulsive forces between the vibrating atoms when they are close to one another and the dissociation when the bond is strongly extended (large displacement).

In the NIR spectrum in addition to overtones also combinations of fundamental tones and overtones are observed. Due to these overtone and combination bands absorption from the different infrared active bonds in molecules often overlap in NIRS. Overtones and combination bands absorb weaker than the fundamental absorption bands which makes it possible to penetrate or even transmit light through solid samples. This in turn makes representative sampling in NIRS easier compared to infrared spectroscopy, but also interpretation more challenging. For more detailed information please refer to Siesler.[63]

Paper II

Monitoring Process Water Quality Using Near Infrared Spectroscopy and Partial Least Squares Regression with Prediction Uncertainty Estimation

Peter B. Skou, Thilo A. Berg, Stina D. Aunsbjerg, Dorrit Thaysen, Morten A. Rasmussen and Frans van den Berg

Applied Spectroscopy, 71, 2017, 410–421.

Article

Applied Spectroscopy

Applied Spectroscopy
2017, Vol. 71(3) 410–421

DOI: 10.1177/0003702816654165
journals.sagepub.com/home/asp
SAGE

Monitoring Process Water Quality Using Near Infrared Spectroscopy and Partial Least Squares Regression with Prediction Uncertainty Estimation

Peter B. Skou[1], Thilo A. Berg[2], Stina D. Aunsbjerg[3], Dorrit Thaysen[4], Morten A. Rasmussen[1,5], and Frans van den Berg[1]

Abstract

Reuse of process water in dairy ingredient production—and food processing in general—opens the possibility for sustainable water regimes. Membrane filtration processes are an attractive source of process water recovery since the technology is already utilized in the dairy industry and its use is expected to grow considerably. At Arla Foods Ingredients (AFI), permeate from a reverse osmosis polisher filtration unit is sought to be reused as process water, replacing the intake of potable water. However, as for all dairy and food producers, the process water quality must be monitored continuously to ensure food safety. In the present investigation we found urea to be the main organic compound, which potentially could represent a microbiological risk. Near infrared spectroscopy (NIRS) in combination with multivariate modeling has a long-standing reputation as a real-time measurement technology in quality assurance. Urea was quantified Using NIRS and partial least squares regression (PLS) in the concentration range 50–200 ppm (RMSEP = 12 ppm, $R^2 = 0.88$) in laboratory settings with potential for on-line application. A drawback of using NIRS together with PLS is that uncertainty estimates are seldom reported but essential to establishing real-time risk assessment. In a multivariate regression setting, sample-specific prediction errors are needed, which complicates the uncertainty estimation. We give a straightforward strategy for implementing an already developed, but seldom used, method for estimating sample-specific prediction uncertainty. We also suggest an improvement. Comparing independent reference analyses with the sample-specific prediction error estimates showed that the method worked on industrial samples when the model was appropriate and unbiased, and was simple to implement.

Keywords

Multivariate calibration, near-infrared spectroscopy, NIRS, real-time monitoring, uncertainty estimation, water quality

Date received: 6 January 2016; revised: 5 May 2016; accepted: 9 May 2016

Introduction

Membrane filtration technologies can be characterized by their ability to separate molecules and constituents of different sizes. Membrane types vary from microfiltration to separate larger particles such as bacteria and fat globules, down to reverse osmosis (RO) membranes that retain nearly everything except water. Membrane systems are implemented in several steps in the dairy production chain to separate molecules of varying sizes in making new products and ingredients. This gentle way of processing dairy products is projected to grow significantly in the coming years.[1]

[1]Spectroscopy and Chemometrics section, University of Copenhagen, Denmark
[2]Dairy, Meat and Plant technology section, University of Copenhagen, Denmark
[3]Food Microbiology section, University of Copenhagen, Denmark
[4]Arla Foods Ingredients, Videbæk, Denmark
[5]Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Corresponding author:
Peter B. Skou, Spectroscopy and Chemometrics section, Department of Food Science, Faculty of Science, University of Copenhagen, Rolighedsvej 26, DK-1958 Frederiksberg C, Denmark.
Email: Peter.b.skou@food.ku.dk

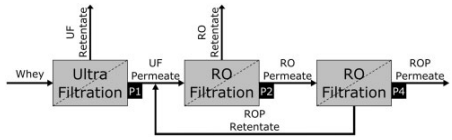

Figure 1. Schematic overview of whey UF processing and RO(P) lactose and water recovery. The three permeate streams are marked by P1, P2 and P4.

At Arla Foods Ingredients (AFI), cheese whey is used as the raw material for new food ingredients. Cheese whey proteins are up-concentrated (retained) by ultra-filtration membranes (UF) while the permeate is further processed by a two-stage RO plant (RO plus RO-polisher; ROP) to collect the lactose and produce ultrapure process water (Figure 1). The cheese whey proteins can be fractionated into a wide range of products with different functional properties. Lactose is dried and can be used, for example, in infant formula or as a food ingredient. Our interest here is the potential for re-using the ROP permeate of this last process step as a water source.

The water footprint of the dairy industry can be greatly reduced by efficient reuse of process water downstream in production flows or cleaning processes. Whey—or, even more upstream, milk—as a raw material consists to a large extent of water, and is the source of process water. However, microbiological safety and hygiene is of utmost importance in the food and dairy industry, and lactose, urea, salts and other organic compounds have been reported to pass RO membranes.[2,3] This indicates that the process water can theoretically act as a growth medium, constituting a potential hazard. As a safety precaution for this potential hazard AFI only keeps the process water for a maximum of 24 h, thus limiting the water reuse strategy. In order to assess the risk of reusing process water, the quality must be continuously guaranteed. The risk can be monitored continuously and in real-time by applying the principles of process analytical technology (PAT)[4] to the process water streams.

In the present study, urea has been identified by GC-MS analysis as the main chemical compound in the RO(P) permeate at AFI. The small organic molecule urea (H_2N–(C=O)–NH_2, 60.06 g mol^{-1}) can be quantified with high accuracy by enzymatic assays in laboratory conditions, but this approach is time-consuming and labor-intensive. Urea can also be detected by near infrared spectroscopy (NIRS),[5,6] which opens up rapid, less labor-intensive or even fully automated measurements. Despite NIRS often being classed as unselective, urea and lactose as interfering species have specific, characteristic absorbance patterns in a region hidden between two dominant water absorption peaks in the spectral range 2080–2325 nm.[6] The urea concentration expected in cheese whey is slightly lower than that which is normally found in bovine milk (~100–600 ppm)[7,8] due to the fact the urea stays in the water phase (whey) during cheese production, and water is added during different cheese manufacturing steps, diluting the whey.

The low urea concentration and the heavy interference with the water bands in NIRS push the detection limit upwards and increases prediction uncertainty. Multivariate modeling techniques such as partial least squares regression (PLS) can lower the prediction error and detection limit compared to univariate methods. However, estimating the prediction uncertainty becomes more complicated. Without a reliable estimate of the prediction uncertainty the detection limit cannot be established and the real-time (microbiological) risk assessment cannot be performed.

Prediction uncertainty estimation for PLS regression has been available for some time in the literature[9] and newer developments have been reported.[10] The quantification of the prediction uncertainty in our work is based on Faber and Kowalski's error-in-variables approach,[9] later modified by Andersen and Bro.[11] Sample-specific prediction uncertainties must be used in multivariate modeling since it cannot be expected that all samples are predicted equally well due to, e.g., differences in concentration of the interfering species. Moving to sample-specific prediction uncertainty presents a clear conceptual difference to the normal model evaluation based on overall or average model performance. Moreover, there is a lack of a proper and straightforward applicable strategy for uncertainty estimation in PLS, despite the theory being developed more than a decade ago. This is not helped by the fact that, unfortunately, justifications for the determination of the parameters that go into the uncertainty equations are seldom given.

These obstacles have resulted in a very limited use of confidence limits for PLS-based predictions in industrial applications. Often several choices can be justified when deciding how to fill in the various terms, depending on: the available data, the calibration strategy and the objective of the uncertainty estimates. For example, when deciding to develop a calibration model for real-time monitoring, several choices have to be made with respect to collecting data. Two clear-cut approaches are to (1) mix stock solutions of the analyte of interest and potential interferents in known concentrations (pure laboratory samples) and measure these directly with the instrumental method (in our case NIRS); or (2) use an independent reference method on grab samples collected from the process and relate these estimates to the instrumental measurements. In the present study we compare these two predominant calibration strategies for the prediction of urea concentration in dairy process water, and present a straightforward strategy for applying and interpreting sample-specific prediction uncertainties. Finally, we demonstrate that NIRS coupled with PLS has the potential for real-time risk assessment of process water.

Theory

The multivariate partial least squares (PLS) calibration model can be described as Eq. 1[12]

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{b}$ is the regression vector relating instrumental measurements $\boldsymbol{X}$ ($N \times J$) (the independent) to the reference values $\boldsymbol{y}$ (one dependent, hence so-called PLS1 is used throughout this work).

The sample-specific variance of the prediction for the model in Eq. 1 is given in Eq. 2[11]

$$s^2_{PE_i} = \overbrace{\left(\frac{1}{N} + h_i\right)\left(s^2_e + s^2_{\Delta y} + \| \boldsymbol{b} \|^2 s^2_{\Delta x}\right)}^{Calibration\ error} + \overbrace{s^2_e}^{Model\ error} + \overbrace{\| \boldsymbol{b} \|^2 s^2_{\Delta x,\,\mathrm{pred}}}^{Error\ in\ X} + \overbrace{s^2_{\Delta y}}^{Error\ in\ y} \tag{2}$$

where h_i is the leverage of the ith test or prediction sample, and N represents the number of samples in the calibration set; s^2_e denotes the variance of the model error; $s^2_{\Delta y}$ denotes the variance of the dependent values (urea reference in our case); $s^2_{\Delta x}$ denotes the average variance across the J variables for the independents (the NIR spectra in our case) in the calibration set; and $s^2_{\Delta x,\,pred}$ denotes the independents' variance for the test or prediction set. This last term, $s^2_{\Delta x,\,pred}$, could for example represent future samples, collected after model calibration, which might exhibit a different uncertainty level. The regression vector, $\boldsymbol{b}$, from Eq. 1 is included as the squared value of its Euclidean norm.[13] The term $1/N$ should be left out if the dependent data, $\boldsymbol{y}$, is not mean centered, but in the remainder of our work we implicitly work with mean centered data.[11]

In order to estimate the spectral variance (the multivariate independents), the variation around the mean for K repeated measurements on the same sample is calculated as described in Eq. 3[11]

$$s^2_{\Delta x_{n,j}} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{X}_{n.,j} - X_{nk,j}\right)^2 \tag{3}$$

where $\bar{X}_{n.,j}$ is the mean value based on K replicated measurements at wavelength j (for sample n), and $X_{nk,j}$ is the kth replicate of sample n at wavelength j. For all combinations of n and j Eq. 3 leads to a matrix of variances of size $N \times J$ (samples $\times$ variables), where a given element contains the variation of a given wavelength for a given sample. It should be noted that all calculations are based on the preprocessed data used for the regression model as indicated in Eq. 1 and not on the raw data. Furthermore, the number of replicated measurements K is assumed constant for all samples; if this is not the case the variance terms should be modified accordingly by weighing with the number of replicated measurements. The variance is then pooled for each wavelength and averaged over all wavelengths resulting in a scalar representation of the spectral variation

$$s^2_{\Delta x} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{n=1}^{N} s^2_{\Delta x_{n,j}} \tag{4}$$

An important step when estimating the spectral variance on the same sample or standard is the choice of the type of replicate measurements. For example, the spectral variance over short-term, consecutive measurements can be used as a *repeatability* uncertainty. The spectral variation can also be determined on standards or samples across several analysis days. In the case of NIRS this might translate into a new background or blank being used and in this case the long-term spectral reproducibility is considered part of the spectral variance.

The *error in y*, the univariate dependent $s^2_{\Delta y}$, can be estimated from replicate measurements similarly to the spectral variation, or for example simply taken from the analytical uncertainty reported by the laboratory or the supplier of an analytical reference method. It can however be assumed that the most realistic estimate will come from estimating the variation under the applicable laboratory conditions, rather than a supplier's reported variation.

The *model error* is given in Eq. 6 as the observed error of the PLS calibration model (mean squared error of calibration (MSEC); Eq. 5), reduced by the *error in y* and the *error in X*

$$\mathrm{MSEC} = \frac{\sum_{n=1}^{N}(\hat{y}_n - y_n)^2}{N - A - 1} \tag{5}$$

$$s^2_e = \mathrm{MSEC} - s^2_{\Delta y} - \| \boldsymbol{b} \|^2 s^2_{\Delta x} \tag{6}$$

where A is the number of components in the PLS model. The *model error* thus contains errors not directly related to independent $\boldsymbol{X}$, dependent $\boldsymbol{y}$ or the calibration. This error includes, e.g., deviations from the Beer–Lambert law as well as other deviations from the bi-linearity assumption such as shifts in the NIRS spectra. Andersen and Bro[11] state that the estimate for MSEC must provide a reasonable estimate of the predictive ability. The model error may be inappropriate if the model is under- or over-fitted, and the estimate of the prediction uncertainty is not valid or in other terms, is biased.

Simplifying the uncertainty estimate can be done by substituting the *model error* (Eq. 6) into Eq. 2 under the assumption the spectra variation in the calibration and prediction samples is similar:[13]

$$s^2_{PE_i} = \left(\frac{1}{N} + h_i\right) \cdot \mathrm{MSEC} + \mathrm{MSEC} \tag{7}$$

The *error in X* term, $\|\boldsymbol{b}\|^2 s^2_{\Delta x}$, assumes homoscedastic measurement errors and no collinearity between the variables in $\boldsymbol{X}$.[11] For NIRS measurements these assumptions do not hold. Spectroscopic data such as NIRS analyses is highly co-linear in the wavelength direction and therefore an alternative formulation is suggested. Rather than multiplying the square of the norm of the regression vector (related to the so-called net analyte signal (NAS) defined by Lorber et al.[14]) with the variation in $\boldsymbol{X}$, the regression vector should be weighed by the size of the variance-covariance matrix of $\boldsymbol{X}$, mean centered per sample

$$\tilde{X}_{nk,j} = X_{nk,j} - \bar{X}_{n.j} \tag{8}$$

where $\bar{X}_{n.j}$ is the mean of variable (e.g., wavelength) j over the K replicates for sample n, $X_{nk,j}$ is the original value for variable j in replicate k for sample n, and $\tilde{X}_{n,j}$ is thus the mean centered sample for variable j. Running over all replicates and samples a matrix $\tilde{\boldsymbol{X}}$ is generated ($N \times J$). From this the variance-covariance matrix is derived

$$\boldsymbol{C} = \frac{1}{NK - N}\left(\tilde{\boldsymbol{X}}^T \cdot \tilde{\boldsymbol{X}}\right) \tag{9}$$

Here, $\boldsymbol{C}$ ($J \times J$) is the variance–covariance matrix (symmetric, with the variance for each wavelength on the diagonal and covariances between two different wavelengths on the off-diagonals). The number of degrees of freedom is the total number of spectra (NK), subtracting the number of independent samples (N) due to the local mean centering operation in Eq. 8. The regression vector can now be weighted by $\boldsymbol{C}$

$$Error\ in\ X = \boldsymbol{b}^{\mathrm{T}} \cdot \boldsymbol{C} \cdot \boldsymbol{b} \tag{10}$$

Here, the alternative *error in X* is also a scalar (in line with a recently suggested approach by Allegrini et al.[15]). Substituting Eq. 10 into Eq. 2 generates an alternative uncertainty estimator

$$s^2_{PE_i} = \left(\frac{1}{N} + h_i\right)\left(s^2_e + s^2_{\Delta y} + \boldsymbol{b}^{\mathrm{T}} \cdot \boldsymbol{C} \cdot \boldsymbol{b}\right) + s^2_e + \boldsymbol{b}^T \cdot \boldsymbol{C}_{\boldsymbol{pred}} \cdot \boldsymbol{b} + s^2_{\Delta y} \tag{11}$$

Substituting Eq. 6 into Eq. 11, the same reduced form as shown into Eq. 7 will be obtained, under the same assumptions concerning calibration and prediction.

Methods and Materials

Process Sample Collection

Process water samples were collected simultaneously from UF, RO, and ROP permeate in a full-scale production system at AFI (40×3 permeate samples; Figure 1). Sampling was initiated immediately after the treatment plant was started (following a cleaning-in-place) and continued for approximately 10 h. This time period included two feed tank changes where samples were collected with high frequency and with low sampling frequency in between (Figure 2). All samples were accumulated in 250 ml amber, sterile, polypropylene sample bottles (ISOLAB, Germany) in less than 1 min, and stored at 5 °C for a maximum of three days before measurements. Immediately before sampling, the process valves were opened and flushed-to-waste for approximately 10 s; conductivity was measured on-site directly after collection for quality assurance purposes (Figure 2).

Independent urea concentration determinations were made on 69 out of 120 samples using a urea + ammonia enzyme assay (Megazymes, UK), measuring absorption at 340 nm with a ThermoScientific Evolution 220 UV/Vis spectrometer as specified by the vendor. The uncertainty ($s_{\Delta y}$) was estimated to be approximately 1 ppm, based on replicate estimates ($K = 5/6$) of random process samples ($N = 4$) and laboratory standards ($N = 2$; bias estimated from the standards only assessed at 1 ppm). Accuracy was found to be in compliance with the laboratory sample label values and the supplier's specifications.

Laboratory Sample Preparation

To mimic the process water samples, standards of urea and lactose were made by combining in varying ratios nine urea stock solutions, three lactose stock solutions, and demineralized water. The aim was to fill out a calibration space representative of the three process water streams (UF, RO, and ROP permeate). Thirty suggested combinations were calculated covering 56–667 ppm urea and 125–4800 ppm lactose. From this set 10 combinations that cover the design space uniformly were selected with the Kennard–Stone algorithm.[16] The 10 combinations were supplemented with the three lowest urea stock solutions containing 46, 110 and 156 ppm urea, as well as one lactose stock containing 6061 ppm. In total, $N = 32$ measurements were performed, spread over three measurement days by selecting a subset of the 14 laboratory samples each measurement day.

Multivariate Calibration

Two separate PLS models were built relating the NIRS spectra to the concentration of urea: (1) a model based on data from laboratory standards made of known amounts of urea and lactose (herein denoted lab model); and (2) a model based on urea concentration determined by the enzyme assay reference method applied to process samples (process model). The process grab samples were split into a calibration set for the process model, and an independent validation set applied to both PLS models. The calibration

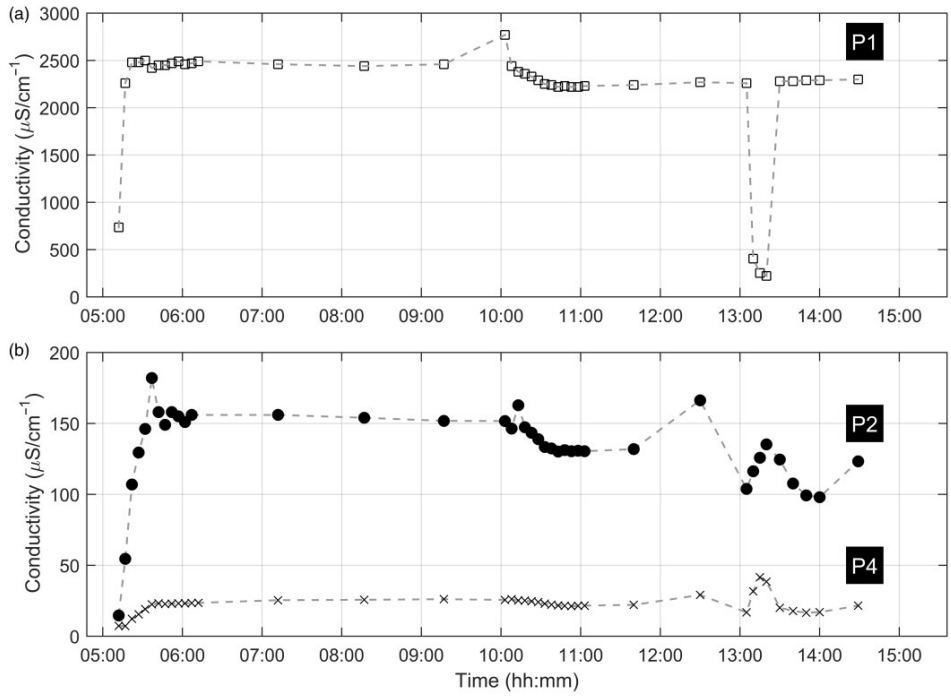

Figure 2. UF and RO(P) sample collection trajectory illustrated by conductivity values; production start 05:12 h, and two feed tank changes took place at 10:05 h and 13:05 h. (a) P1 = ultrafiltration permeate; (b) P2 = reverse osmosis permeate, P4 = reverse osmosis polisher permeate.

set for the process model was the startup of the RO plant from 05:12 to 09:17 h (Figure 2, $N = 33$ samples with urea estimates from the three permeate sample collection locations), while the common validation set was formed by samples taken from 10:03 to 14:29 h ($I = 12 \times 3$ samples). One sample was removed from the process model calibration data as the urea concentration estimate was outside the linear range of the enzyme assay (RO permeate 05:22 h, $N = 32$). The preprocessed data were mean centered before fitting PLS models. Model complexity during calibration was estimated by inspecting the shape of the regression vector, reduction in root mean square error of calibration (RMSEC) and root mean square error of cross validation (RMSEC) and the difference between the aforementioned values. Partial least squares regression models were built using the PLS toolbox (Ver. 7.9.5, Eigenvector Research) in Matlab 2015a (MathWorks, Inc.). Prediction uncertainty estimates were calculated in Matlab using in-house routines based on Eq. 2 and Eq. 11.

Near-Infrared Spectroscopy

Near infrared spectroscopy spectra were measured with an ABB Bomem MB Series FT-NIRS (Canada) with a custom-made, temperature-controlled sample flow cell with a path length of approximately 1 mm. Each sample was introduced into the cell and measured five times in a row, each the average of 128 scans (total duration approximately 6 min) over the spectral range 14 285–4000 cm^{-1} or 700–2500 nm with a spectral resolution of 8 cm^{-1}. The temperature of the flow cell was set to 27 °C, and we assessed that the maximal temperature differences between the five replicates was 0.1 °C. All samples were measured over the course of three days, and the same background – air, empty cell plus a 25% absorbance optical filter – was used for all spectra obtained. For urea calibration the spectral range 2083–2257 nm was selected, and the spectra were preprocessed using a Savitzky–Golay second order polynomial fitting, second derivative, with a window size of 17 points (approximately 62 nm).[17] For all samples, four of the five measured spectra were excluded based on the Euclidean norm of the difference between the average of all five preprocessed spectra and each given preprocessed spectra. The spectrum closest to the mean was included in the calibration data set with its corresponding urea reference value. The spectral variation was estimated in two ways: (1) from the three most similar replicate measurements (discarding two out of five rather than four

spectra in the spectral selection procedure described above) for each sample obtained for the process model data set; (2) from the variation between three replicates of those standards measured on all three measurement days (only available for the lab model data set).

Results and Discussion

Regression Modeling

The present work describes two clear-cut approaches for building prediction models for process monitoring. One approach is to prepare laboratory samples reflecting the expected variation in the process of interest. The advantage is that the calibration space can be spanned by a small number of samples of known concentration, making it fast and economic. The risk is that the chemical composition of the process is not captured, for example, by missing minor interferents. The other approach is collecting actual process samples and performing independent reference analysis, making full use of the inverse-calibration concept behind PLS.[12] The advantage is that the process composition is incorporated in the independents, assuming that the right samples and a sufficient number of samples are collected. Reference analysis is, however, often time-consuming and expensive since typically many samples are needed to span the calibration space using grab sampling plus chance.

Depending on which calibration building approach is taken different inputs are available for estimating the *error in X*. Using laboratory standards, the same sample can be measured repeatedly over time to incorporate long-term reproducibility. However, a risk is that the determined repeatability or reproducibility does not represent the actual process sample composition, making the estimate inapplicable. Alternatively, using process samples the same sample material is not available on different days, forcing the use of short-term repeatability based on repeated consecutive measurements. The two approaches will also lead to models with a different predictive performance and uncertainty.

Figure 3 shows in two ways the average spectrum of all lab model NIR spectra: with air and demineralized water as reference. The shaded area shows the calibration range applied for both PLS models and illustrates that the shoulder of the left-hand water band, a NIR combination band, is included in the calibration range. It is well known from the literature that small temperature changes shift the NIR spectra's water bands left and right.[18] These temperature shifts resulted in an uncontrolled sign-change of the left-hand tail of a preprocessed spectrum if the temperature was even marginally lower or higher than the background measurement when using demineralized water as a blank. Air was used as the background to avoid this issue. To minimize the temperature/shift-induced variation from the

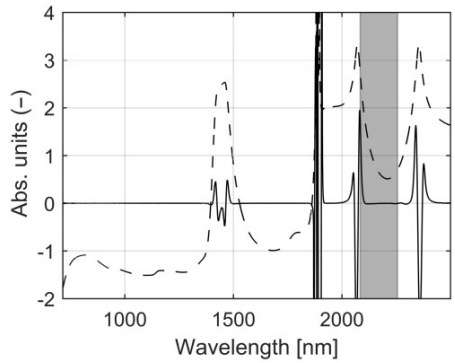

Figure 3. Average spectra of lab model data set with air (solid) and water (dashed, intensity-scaled for visual comparison) as background; shade area is the range used in urea PLS regression modeling.

PLS models the spectrum closest to the average of all the five spectra was selected. The objective was to avoid the more *extreme* spectra by choosing the most representative one. However, it cannot be guaranteed that the average temperature of the spectra obtained from each samples are identical, and small temperature-induced shifts will still be present. Interestingly, excluding the highly temperature-sensitive water band shoulder from the calibration range was not an option since this resulted in a lowered predictive performance from the models. Including more of the lactose absorption range above 2250 nm (data not shown) also resulted in reduced predictive performance.

Figure 4a–c show the preprocessed and mean centered calibration data for the two PLS models, while Figure 4d shows the corresponding regression vectors. The extent of the temperature shift can be observed in Figure 4b where the five consecutive measurements from one lab model sample (high urea, medium lactose) are shown together with the spectral variation of the full calibration data set. The selected spectrum for this sample can also be seen in Figure 4a, together with a low urea–high lactose sample and a medium urea–low lactose sample. The urea signal can be identified in Figure 4b in the high urea–medium lactose sample, where high urea concentration results in two valleys at 2150 nm and 2200 nm and a peak around 2170 nm in the preprocessed (second derivative) spectrum, in agreement with Shaw et al.[6] and Olesberg et al.[5] Lactose absorbance can be seen very clearly in Figure 4c where UF permeate samples contain high amounts of lactose, while RO permeate and ROP permeate samples do not. High lactose manifests itself as high absorbance around 2240 nm in the preprocessed spectrum in accordance with Olesberg et al.[5]

Two models were built (Figure 5 and Table I): (1) on the laboratory samples to yield the lab model, and (2) a model

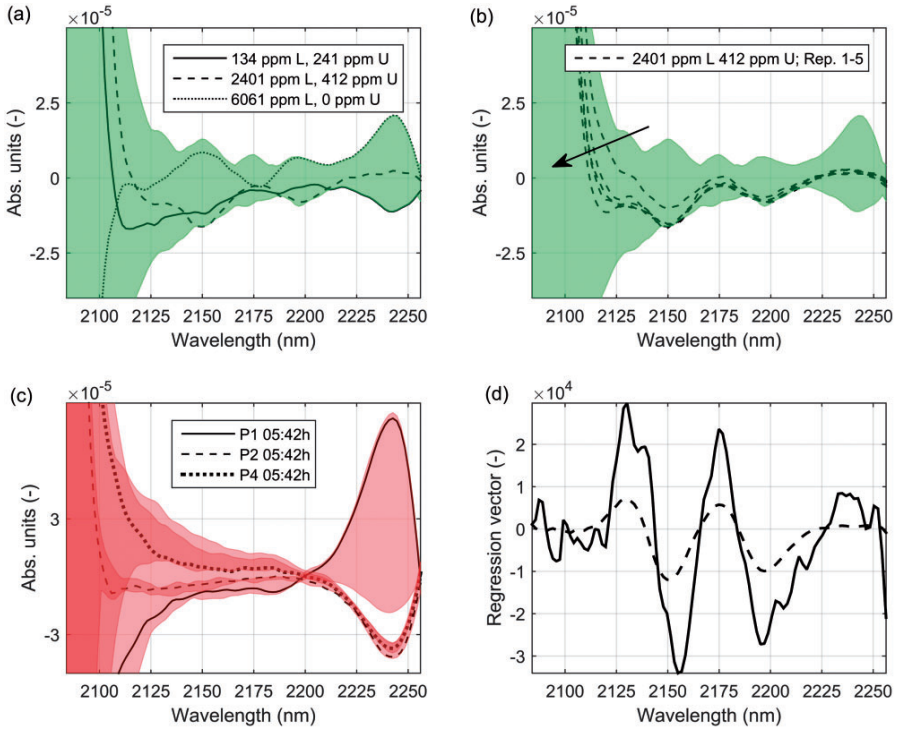

Figure 4. NIR spectra of (a) lab model data (shaded area is convex envelope) with three selected samples; (b) lab model data with one sample in five consecutive measurements; arrow illustrating shift/order over time; (c) process model data with three samples from the same time point (05:42 h) from different sampling locations–spectral variation from each location shaded convex envelope; (d) PLS regression vectors for lab model (dashed) and process model (solid).

built on the process samples to yield the process model. The complexity of both PLS models was estimated to be five components. The two first components in the lab model describe mainly ***X***-variance (99.0% and 0.8% explained variance) and almost no variation in **y**-variance (0.3% and 15.7%). For the process model the first two components explain relative more variation in **y** (17.8% and 50.4%), but still primarily focusing on explaining ***X*** (91.7% and 7.8%). This indicates that both data sets have large amounts of spectral variation not related to the urea information, and preprocessing has not been successful in bringing the relevant information forward. Inspecting the regression vectors in Figure 4d, the process model appears over-fitted based on its irregular/non-smooth appearance, also resulting in a large norm as seen in Table I. The process model's calibration (Figure 5c and Table I) and prediction performance (Figure 5d and Table 2) are close to each other, and in fact appear not to be over-fitted (albeit with a bias in the high-range/P1 validation results). The calibration and validation performance for the lab model can be seen in Figure 5a and 5b and in Tables I and 2. The urea calibration range is far wider than the validation samples' concentration, and the root mean square error of prediction (RMSEP) increases considerably compared to the RMSEC (13.0 ppm versus 9.7 ppm). Especially, UF permeate/P1 samples are predicted with large errors. Inspecting the spectral variation of the lab model samples in Figure 4b and comparing with the process samples' variation in Figure 4c, it is clear that the lab model samples did not span a wide enough lactose variation and the urea range span was too wide.

Estimating Prediction Uncertainty

The prediction error variance was calculated according to Eq. 2 and Eq. 11 using the values presented in Table 3. Since the prediction uncertainty is sample specific, each variance component was extracted and averaged over the calibration and test set, before presentation in the table. The variance of the prediction error (s^2_{PE}) was converted to standard deviation of the prediction error (s_{PE}) by taking the square root. These values are tabulated together with

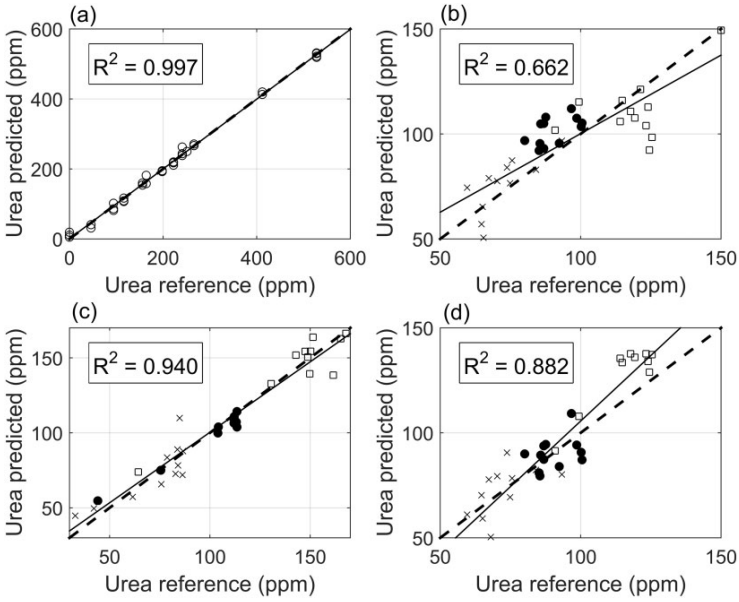

Figure 5. Partial least squares regression models for urea; (a) lab model calibration, (b) lab model validation, (c) process model calibration, (d) process model validation. Symbols used: P1 samples, open squares; P2 samples, full circles; P4 samples, x; Lab samples, open circles; x = y fit, dashed line; best fit through data, solid line.

Table 1. Summary of PLS calibration models.

Model	*A*	*N*	h_n	$\|\|b\|\|$	RMSEC/MSEC (ppm/ppm^2)
Lab	5	32	0.05–0.50	4.81×10^4	9.7/95
Process	5	32	0.06–0.31	144×10^4	10.2/103

Table 2. Summary of validation of PLS models on the test set.

Model	*I*	h_i	RMSEP (ppm)
Lab	36	0.05–10.37	13.0
Process	36	0.03–2.24	12.1

the prediction performance (RMSEP) of the validation set for comparison.

When comparing the average prediction performance (RMSEP) with the averaged estimated prediction uncertainty (s_{PE}) the difference is small for the process model, while the lab model shows a larger discrepancy. An explicit bias term is not part of the prediction uncertainty error and this might be the reason for this discrepancy as the lab model exhibited some bias when validated (Figure 5b). Regardless of which formulation of the uncertainty estimate is applied the estimate will be the same as both Eq. 2 and Eq. 11 reduce into Eq. 7 when spectral variance during prediction and calibration is assumed to be the same. Such insight into model predictive performance can prove valuable, and this example serves as evidence that the estimates are reasonable, trustworthy and easy to obtain.

The *error in y* was set to 1 ppm since this was both the uncertainty claimed and the uncertainty found through repeated measurements. Using the claimed uncertainty of the vendor should be done with caution as this may not reflect the uncertainty achieved under the given laboratory conditions.

The calibration error is determined by the size of the calibration set (*N*), the prediction sample leverage (h_i), and the calibration uncertainty (MSEC; via Eq. 7), leading to identical variances for both prediction uncertainty formulations (Eq. 2 or Eq. 11). The leverage ranges of the calibration and predicted samples are presented in Tables 1 and 2, respectively. Here, the leverages of the calibration samples are similar for both models, but the leverages of the validation samples are very different. The difference between the two models is explained by the lab model not spanning the validation sample space sufficiently, subsequently leading to larger prediction uncertainties. Intuitively this can be understood from classical univariate calibration where samples at the extremes of the model have larger prediction errors, i.e., prediction uncertainty intervals expand when moving towards the extremes of the

Table 3. Summary of variance components calculated for the two PLS models for Eq. 2 and Eq. 11 based on the test set.

Equation	Model	*Calibration error** (ppm^2)	*Model error*/s^2_e (ppm^2)	*Error in X* (ppm^2)	*Error in y*/$s^2_{\Delta y}$ (ppm^2)	S_{PE}*/S^2_{PE}* (ppm/ppm^2)
2						
	Lab	192	89	3.6	1	15.1/286
	Process	53	52	51	1	12.4/157
11						
	Lab	192	93	0.1	1	15.1/286
	Process	53	99	3.6	1	12.4/157

*Based on Eq. 6 and Eq. 7, the error does not change between the two prediction uncertainties equations.

prediction model's range. It should be recalled that leverage only represents the modeled directions in the data and does not reveal samples not fitting the model. This information is found from the residuals, which should be monitored independently to detect samples not compatible with the model. Hence, this issue is not covered by the uncertainty estimation procedure evaluated here.

The *model error* is directly connected with the estimate of *error in X*. Using Eq. 6 to calculate the *model error*, it becomes larger for the lab model than for the process model, which is a consequence of the achieved MSEC in combination with a low *error in X* and low *error in y*. Intuitively, the *model error* can be understood as the error in the relation between the spectral data, $\boldsymbol{X}$, and the reference data, $\boldsymbol{y}$, that is, deviations from Beer–Lambert law, non-linearity at a very low or high absorbance and violations of the bi-linearity assumption in PLS. In other words, if the error does not originate from spectral data or reference data it must be the lack of information in the spectral data needed to explain the reference values.

In Table 3 the *error in X* when using Eq. 2 is high for the process model and lower for the lab model, despite the process data having a lower spectral variation ($s^2_{\Delta x}$). This intuitive discrepancy is explained by the irregular regression vector seen in Figure 4d, also reflected in the size of the Euclidian norm of $\boldsymbol{b}$ in Table 1. However, since the spectral variance is not homoscedastic (equal variance for all wavelengths), taking the mean of all wavelengths and multiplying this with the squared norm of the regression vector is not satisfactory. Instead, the spectral error, or more precisely the variance-covariance of the spectral error, should be weighted by the size of regression vector at each wavelength. Thus, if a wavelength with high variation (per sample) does not contribute to the prediction (low b_j entry) the propagated uncertainty should be low. From Table 3 the alternative *error in X* (Eq. 11) becomes much lower compared to the original (Eq. 2). Still, the process model shows more *error in X* compared to the lab model, probably due to the noisy regression vector. The unexplained prediction variance released from the alternative *error in X* term is absorbed by the *model error* term instead as discussed above. This redistribution of variance is reasonable since the uncertainty stems not from irreproducible spectra, but, e.g., from temperature-induced shifts. Interestingly, the *model error* for both calibration strategies converge (93 ppm versus 99 ppm), suggesting that the temperature-induced shifts are similar for the two calibration data sets.

It should be noted that the original *error in X* definition ($\|\boldsymbol{b}\|^2 s^2_{\Delta x}$) may under different circumstances underestimate the spectral uncertainty (as seen, e.g., in Allegrini et al.[15]), where it overestimates in our case. It is also important to notice that the *error in X* term only reflects the spectral variation put into the equation, i.e., reproducibility (long-term) or repeatability (short-term). In our particular case, however, it does not make a noticeable difference if samples measured on the same day or measured over three days are used (results not shown). This might be very different if an NIRS-based urea monitoring scheme is implemented over a longer time period, for example months or years. If the calibration model is based on laboratory samples and used in a production environment (e.g., an optical probe mounted directly inside the process, including all the production noise), attention should be turned towards estimating a distinct $\boldsymbol{C}_{pred}$ as the variance–covariance of the calibration data no longer represents that of the predicted samples, and a correction for this is required.

The averaged standard deviations of the prediction uncertainties (s_{PE}) can be directly compared to the RMSEP values. In Table 3, the estimates can be seen to be very close to the observed average prediction error. This suggests that the reduced expression of the uncertainty estimate (Eq. 7) can estimate the predictive performance sufficiently accurate, given that the PLS model is reasonable. In essence the estimate is determined by two terms: (1) the leverage of a sample; and (2) the MSEC. Thus, confidence interval-based limits can be built directly from the PLS model if the partition of the prediction error variation into the individual parts is unnecessary. Conversely, confidence *bands* cannot be built for PLS models with more than

one component. This follows the argument for using sample-specific prediction intervals. In classical univariate regression the bands are a function of the nominal difference, or distance, of the obtained value to the center of the model. Equation 7 can be rewritten as $s^2_{PE_i} = (1 + 1/N + h_i) \cdot \mathrm{MSEC}$ making the connection to classical univariate regression even more obvious. In multivariate regression this distance is given by the leverage. The leverage, however, is determined by the score value for each loading,[19] which means that a sample in the center of the multivariate regression model can have a high leverage stemming from other spectral contributions not related to the analyte of interest. Partial least squares regression models describe other sources of variation not related to the analyte of interest, therefore, the leverage cannot be used in the same way for PLS models with more than one component. For a one-component PLS model the leverage will also be the score to the loading describing the analyte of interest, and only in this case can calibration model confidence bands be used. Instead sample-specific prediction intervals can be calculated: $PI_\alpha = \hat{y}_i \pm t_{\alpha/2,N-A-1} s^2_{PE_i}$, where $t_{\alpha/2,N-A-1}$ denotes the critical value of a *t*-distribution with α-level confidence and $N-A-1$ degrees of freedom.[20]

Process Monitoring

Figure 6 shows predicted urea concentrations together with prediction intervals and the available reference values, with the critical *t*-value of 2.06 for 95% prediction (or confidence) interval with 26 degrees of freedom. For P1 validation samples, the prediction intervals often do not capture the observed prediction uncertainty. Looking back to the scatterplot of the validation data set, it can be seen that P1 samples exhibit a bias. As mentioned above, the uncertainty estimates are only valid when no bias is present and consequently do not work for P1 samples.

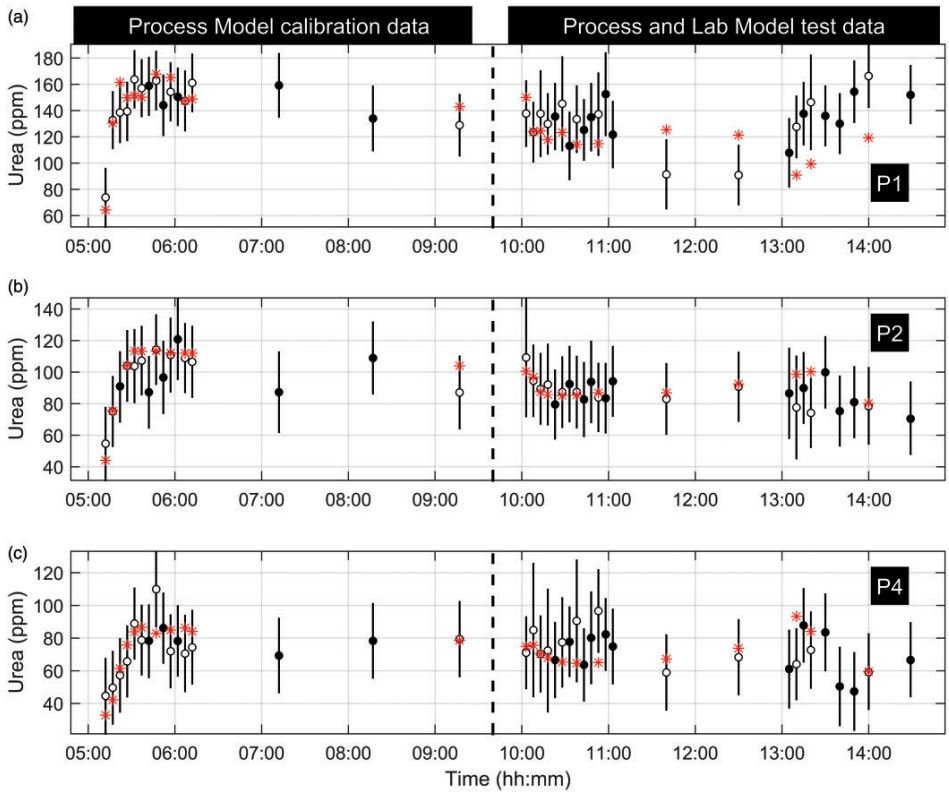

Figure 6. Process model predictions of all NIRS samples with reference values (predictions, open circles; reference values, red stars) and samples without reference samples (closed circles). Vertical lines show the sample specific 95% prediction interval. (a) P1; (b) P2; (c) P4.

A systematic decline in urea concentration is seen from UF permeate to RO permeate to ROP permeate (Figure 6). As expected, urea is not retained completely. Lee and Lueptow[3] reported the same phenomenon for laboratory-scale RO units and hypothesized that urea is very difficult to retain due to the chemical's affinity to the membrane material. In effect, urea seems to reach equilibrium across the membrane(s). If information about the organic load is desired then urea may not be the obvious target molecule. However, urea may in fact be the only organic molecule permeating the membrane system during normal operating conditions (NOC), an observation supported by the very low conductivity values for ROP permeate (Figure 2).

The samples collected in the present study cover the first half of a production cycle and three types of samples. In Figure 6 we observe that the UF permeate samples are predicted with larger uncertainty, while the RO(P) samples show a more consistent picture. During the sample collection run the process startup and two feed tank changes were included. These operations were targeted to observe potential abnormal operating conditions (AOC) and thereby larger variation in the collected samples. However, all process parameters monitored at the production site as well as the laboratory reference measurements and PLS prediction models indicated that the process was in NOC during the tank switches. The dip in conductivity of UF permeate seen in Figure 2 arises from short-time recirculation of permeate into the feed. The only large variation observed was during startup. Since no AOC situation is included in the present study the predictive performance of the models is not truly challenged.

Near-infared spectroscopy can be used to quantify a wide range of different vibrational bonds and could potentially be used for RO performance monitoring via alternative routes. If, for example, leaks appear in the membrane cartridges then some lactose would be expected to seep into the RO and eventually the ROP permeate – a clear AOC situation. Using the methods described in this manuscript calibration models with sample-specific uncertainty, limits can be built. This will, however, not be feasible using a standard NOC process model data set (since the leakage problem is an exception and data collection thus unfeasible). Instead this has to rely on a lab model during the calibration stage and process data during monitoring/prediction.

When building predictive models for process performance monitoring, most often both laboratory samples and process samples are combined to yield the best prediction performance.[21] In the present case the laboratory samples did not span the desired variation and the process samples do not cover large enough process variations from only 10 h of production. Combining the two data blocks did not significantly change the model performance (data not shown). Improved prediction performance would be achieved if laboratory samples were redone and additional process samples collected over a long period of time.

Conclusion

In order to reuse process water in the dairy industry the quality must be guaranteed at all times. Urea may not be the obvious chemical compound for monitoring membrane filtration process performance, but it does constitute the main organic matter. In this work it has been shown that very low concentrations of urea can be quantified by NIRS spectroscopy in connection with PLS regression. Critical evaluation of the PLS predictions can only be trusted if the prediction uncertainty is quantified. From the prediction uncertainty estimation, diagnostics can be extracted for evaluation of the model adequacy (calibration error) as well as evaluation of the measurement setup (*error in X*). This work shows that prediction uncertainties for PLS regression models are not complicated to obtain, regardless of which calibration strategy is used, under the condition that the model is appropriate and unbiased.

Conflict of Interest

The authors report there are no conflicts of interest.

Funding

This work was supported by the REuse of WAteR in the food and bioprocessing inDustries (REWARD) consortium funded by the Danish council for Strategic Research, Programme Commision on Health Food and Welfare.

References

1. H.H. Holst, F. van den Berg, S.B. Engelsen. "Water Reuse and Saving in the Food Industry: A New Frontier in Food Manufacturing". New Food. 2014. 17: 7–11.
2. M. Vourch, B. Balannec, B. Chaufer, G. Dorange. "Treatment of Dairy Industry Wastewater by Reverse Osmosis for Water Reuse". Desalination. 2008. 219: 190–202.
3. S. Lee, R.M. Lueptow. "Reverse Osmosis Filtration for Space Mission Wastewater: Membrane Properties and Operating Conditions". J. Membr. Sci. 2001. 182: 77–90.
4. F. van den Berg, C.B. Lyndgaard, K.M. Sørensen, S.B. Engelsen. "Process Analytical Technology in the Food Industry". Trends Food Sci. Technol. 2013. 31: 27–35.
5. J.T. Olesberg, M.A. Arnold, M.J. Flanigan. "Online Measurement of Urea Concentration in Spent Dialysate during Hemodialysis". Clin. Chem. 2004. 50: 175–181.
6. R.A. Shaw, S. Kotowich, H.H. Mantsch, M. Leroux. "Quantitation of Protein, Creatinine, and Urea in Urine by Near-Infrared Spectroscopy". Clin. Biochem. 1996. 29: 11–19.
7. J.S. Jonker, R.A. Kohn, R.A. Erdman. "Using Milk Urea Nitrogen to Predict Nitrogen Excretion and Utilization Efficiency in Lactating Dairy Cows". J. Dairy Sci. 1998. 81: 2681–2692.
8. M. Aguilar, M.D. Hanigan, H.A. Tucker, B.L. Jones, S.K. Garbade, M.L. McGilliard, C.C. Stallings, K.F. Knowlton, R.E. James. "Cow and Herd Variation in Milk Urea Nitrogen Concentrations in Lactating Dairy Cattle". J. Dairy Sci. 2012. 95: 7261–7268.

9. N.M. Faber, B.R. Kowalski. "Prediction Error in Least Squares Regression: Further Critique on the Deviation used in The Unscrambler". Chemom. Intell. Lab. Syst. 1996. 34: 283–292.
10. L. Zhang, S. Garcia-Munoz. "A Comparison of Different Methods to Estimate Prediction Uncertainty Using Partial Least Squares (PLS): A Practitioner's Perspective". Chemom. Intell. Lab. Syst. 2009. 97: 152–158.
11. C.M. Andersen, R. Bro. "Quantification and Handling of Sampling Errors in Instrumental Measurements: A Case Study". Chemom. Intell. Lab. Syst. 2004. 72: 43–50.
12. S. Wold, M. Sjöström, L. Eriksson. "PLS-regression: A Basic Tool of Chemometrics". Chemom. Intell. Lab. Syst. 2001. 58: 109–130.
13. N.M. Faber, R. Bro. "Standard Error of Prediction for Multiway PLS 1. Background and a Simulation Study". Chemom. Intell. Lab. Syst. 2002. 61: 133–149.
14. A. Lorber, N.M. Faber, B.R. Kowalski. "Net Analyte Signal Calculation in Multivariate Calibration". Anal. Chem. 1997. 69: 1620–1626.
15. F. Allegrini, P.D. Wentzell, A.C. Olivieri. "Generalized Error-Dependent Prediction Uncertainty in Multivariate Calibration". Anal. Chim. Acta. 2016. 903: 51–60.
16. R.W. Kennard, L.A. Stone. Technometrics. 1969. 11: pp 137–148.
17. Å. Rinnan, F. van den Berg, S.B. Engelsen. "Review of the Most Common Pre-Processing Techniques for Near-Infrared Spectra". TrAC, Trends Anal. Chem. 2009. 28: 1201–1222.
18. F. Wülfert, W.T. Kok, A.K. Smilde. "Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models". Anal. Chem. 1998. 70: 1761–1767.
19. D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke. "Multiple and Polynomial Regression". In: Handbook of Chemometrics and Qualimetrics, Part A. Amsterdam: Elsevier, 1997. Pp. 263–303.
20. N.M. Faber, D.L. Duewer, S.J. Choquette, T.L. Green, S.N. Chesler. "Characterizing the Uncertainty in Near-Infrared Spectroscopic Prediction of Mixed-Oxygenate Concentrations in Gasoline: Sample-Specific Prediction Intervals". Anal. Chem. 1998. 70: 2972–2982.
21. C.B. Zachariassen, J. Larsen, F. 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. 2005. 76: 149–161.

3.1 Estimating the prediction error — continued

As discussed in **Paper II** the expression for estimating the sample-specific prediction error can be reduced to

$$\widehat{s^2}_{PE_i} = \left(1 + \frac{1}{N} + h_i\right) \cdot MSEC$$

Intuitively this is a satisfying formulation when one considers that it consists of three distinct parts: 1, the calibration error which can be considered the minimal error one can expect from a normal sample; 2, one over the number of samples $(1/N)$ in the calibration set, which represents the robustness of the calibration and shrinks for a larger dataset; and 3, the distance to the center of the PLS model for this sample, the leverage h_i. It also illustrates that correctly estimating the mean square error of calibration, MSEC, turns out to be the most important figure of merit, particularly when estimating all the specific error terms is not of interest. The mean square error of calibration is often expressed as

$$MSEC = \frac{\sum_{n=1}^{N} (\hat{y}_n - y_n)^2}{N - A - 1} \tag{3.1}$$

In Equation 3.1 $N - A - 1$ represents the degrees of freedom. However, this only holds for true linear models such as multiple linear regression, where A represents the number of response variables. With PLS (a bilinear model) A is the number of components. However, it has been found that one PLS component consumes more than one degree of freedom,[64] making Equation 3.1 optimistic.

In the case where the number of calibration samples is low (small N) and a complex model is needed (large A), the MSEC becomes very sensitive to the choice of the denominator. The prediction error also inflates with a low number calibration samples. In Figure 3.3 the prediction error estimates are shown using $N - 1$ (an optimistic denominator often used in literature) and $N - A - 1$ in the calculation of the calibration error. It's worth mentioning again that the estimated prediction errors are only valid if MSEC is unbias, i.e. the model is not over- or underfitting.

In short, an incorrect choice of MSEC formulation and / or overfitting of the PLS model can have a huge effect on the perceived prediction errors.

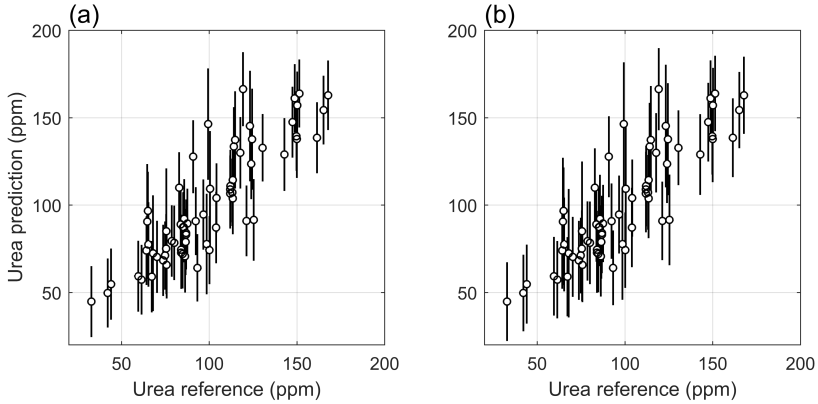

Figure 3.3: Scatterplot of the urea concentration references against the urea concentrations estimated with a five component PLS model. 95% confidence bands estimated with: MSEC estimated with optimistic denominator, $N-1$ (a) compared to $N-A-1$ (b).

3.2 The multivariate detection limit

Closely linked to the estimation of prediction uncertainty in PLS models is the detection limit. For univariate models where the detector signal is specific for a given analyte, the detection limit (DL) is defined by IUPAC[65] as

$$DL = \overline{x}_{bl} + (t_{1-\alpha,\nu} + t_{1-\beta,\nu}) \cdot s_{bl} \tag{3.2}$$

Here $\overline{x}_{bl}$ is the mean of the signals obtained from a blank sample, α and β specify the accepted Type I (false positive) and Type II (false negtive) error rates and s_{bl} the standard deviation of the signals from a measured blank sample. IUPAC recommends that $\alpha = \beta = 0.05$, which makes it possible to reduce Equation 3.2 slightly. This works well for univariate measurements where the signal only depends on the compound of interest, i.e. the signal is specific. This is seldom the case in multivariate calibration, where the signal can also depend on interfering compounds.

In the pursuit to establish figures of merit for the so-called inverse multivariate models (which includes PLS regression models) similar to those for univariate models, Lorber et al.[66] defined the *net analyte signal* (NAS). The NAS seeks to find the part of the analyte signal that is independent from (orthogonal to, or lies in the null space of) the interferents as illustrated in Figure 3.4. Ideally this vector represents the unique signal of the compound of interest, and as a consequence the length can be used for quantification purposes.

The NAS for a sample, $x^*_{k,i}$, is then obtained by projecting it onto the unique analyte space as shown in Figure 3.4. In PLS the regression vector is assumed to

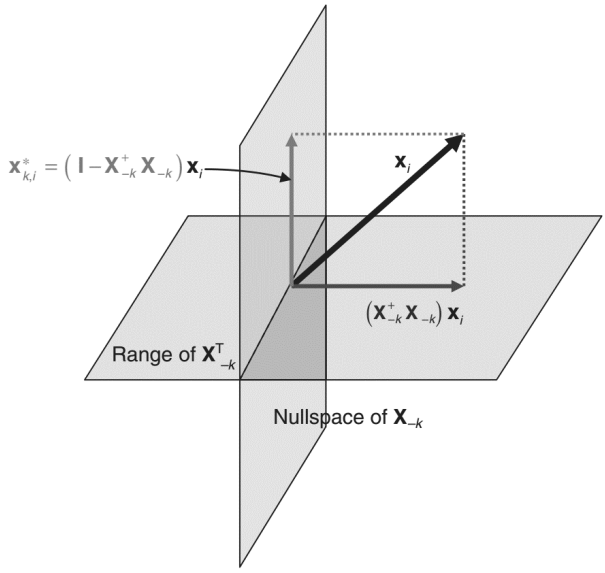

Figure 3.4: Schematic illustration of the net analyte signal projection ($x^*_{k,i}$) into the null space of the interferents ($\boldsymbol{X}_{-k}$) from Kalivas.[68]

point in the same direction as the NAS and with this insight Bro and Andersen[67] show that the NAS can be obtained through the regression vector, $\boldsymbol{b}$:

$$x^*_{k,i} = \boldsymbol{b}\left(\boldsymbol{b}^T\boldsymbol{b}\right)^{-1}\boldsymbol{b}^T\boldsymbol{x}_i$$

The merit of the NAS is that the calibration problem now becomes univariate and figures of merit from univariate statistics (hence Equation 3.2) can be used in describing a multivariate model. While compressing the problem into a univariate problem can seem appealing, it does not follow the argumentation for estimating the prediction uncertainty in **Paper II** requiring *sample-specific* estimates. The argumentation for sample-specific uncertainties for PLS calibrations is still that the uncertainty depends on the concentration of interferents as well as the analyte of interest, or in as stated in **Paper II**, that the distance to the center of the model does not depend on the analyte concentration alone. This is illustrated in Figure 3.5 where the estimated confidence limits per sample is plotted against the estimated urea concentration. The consequence is that every sample will have an associated *multivariate detection limit.*[69] In this framework Boqué et al.[69] as well as Allegrini et al.[70] describe methodologies for obtaining sample-specific multivariate detection limits.

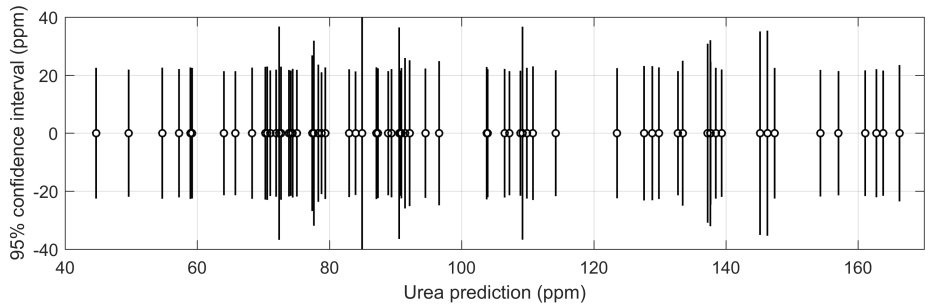

Figure 3.5: Sample-specific prediction errors against prediction urea concentration.

3.3 Tensorisation of NIRS data

The NIR spectrum of water is sensitive to changes in hydrogen-bond strength and hydration.[71,72] Solvents dissolved in water such as ions, organic monomers and polymers can interact with water absorption bands and induce spectral changes. In addition, temperature changes also effect the water spectrum by shifting band positions or absorbance intensities. These changes were identified as the main source of the prediction uncertainty in **Paper II**.

As a consequence, a large number of samples spanning different variations of interferences are needed to make a PLS regression model robust. This is a known challenge in PLS calibration. While it is a burden to collect a relative large data set with the desired variation, the ability to detect unknown interfering compounds (e.g. through residuals) is known as the *first-order* advantage. The term first-order refers to the tensor algebraic notation where a first-order tensor is a vector. This terminology is often used to classify calibrations in terms of the dimensionality of the data.[73–75] The simplest is a zeroth-order calibration where a single response is collected per sample (e.g. urea quantification with enzyme kit), for a first-order calibration a vector is collected per sample (e.g. a NIR spectrum) and when a matrix is collected per sample a second-order calibration may be possible. If a second-order calibration method can be used to quantify urea, a calibration could ideally be constructed from one calibration sample — this is known as the *second-order* advantage. It is therefore highly desirable to establish second-order methods to reduce and simplify the modelling step.

Since temperature can shift the absorption bands in NIR spectra it can be speculated that obtaining NIR spectra a different temperatures can be used to achieve a second-order calibration. Peinado et al.[76] tried this approach on two data sets: 1, spectra of mixtures of water, ethanol and iso-propanol measured at 30, 40, 50, 60 and 70°C in the spectral range from 580 to 1090 nm, and 2, an in-line data

set where six batches consisting of either water, glycerine and ethanol separately or binary mixes were measured during temperature cycles while also diluting the batch. In the study by Peinado et al.[76] PARAFAC was used to decompose the NIR data set in to a trilinear model consisting of scores and two loadings (see Box 4). The obtained scores were related to the reference values (mass fractions of the constituents) by multiple linear regression and the loadings were used to predict scores of the validation samples. Then the multiple linear regression model was used to predict the concentration of the new validation samples from the predicted scores. The purpose of using PARAFAC was to capture the systematic temperature variation in one of the loadings and thereby making the obtained chemical compound scores (representing the concentrations) independent of temperature. Using PARAFAC combined with multiple linear regression it was possible to improve the predictions and obtain more parsimonious models compared to PLS regression models. The predictions were especially improved when the test set was measured at a temperature which was different from the calibration set (as inter- or extrapolation). In an attempt to counter the temperature variation found in **Paper II**, an experiment was conducted to create such a three-way NIR data set.

My experimental set-up consisted of a peristaltic pump that circulated the sample from the reservoir across the flow-cell and back to the reservoir as shown in Figure 3.7. The samples were gradually heated by a magnetic stirrer heating plate and the actual temperature of the solution was measured with a Pt100 probe and logged with a pHTemp2000 pH and temperature data logger (MadgeTech, U.S.). The NIR spectra were acquired every 30 seconds during heating of the samples on the same instrument and settings as described in the **Paper II**. The clock on both the NIR spectrometer and the datalogger were synchronised and NIR spectra were matched up with temperatures based on time of acquisition. Linear interpolation between the obtained temperatures was used to compile spectra from 30-50°C with 1°C intervals. Thus, every sample consisted of 21 NIR spectra obtained at 30-50°C for each sample. NIR spectra were pre-processed using Savitzy-Golay second derivative, after which the data cube was assembled such that samples constituted the first mode, temperature the second mode, and wavelengths the third mode. The dataset was then mean centered on the first mode (the sample mode). The seven samples were either urea or lactose in water, or mixtures of these two compounds as shown in Table 3.1.

Box 4: PARAllel FACtor analysis - PARAFAC

PARAFAC is a multi-way decomposition method that can be viewed as a generalisation of the two-way method pricipal component analysis.[77,78] For simplicity and the fact that PARAFAC is often used to solve trilinear problems, it is here explained for the three-way case only. For a three-way data set, $\underline{\boldsymbol{X}}$ (IxJxK), a three-way PARAFAC model is given by three loading matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ where single elements can be expressed as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$

Loading matrix $\boldsymbol{A}$ is referred to as *scores* and it contains the sample specific information (usually concentrations). Graphically PARAFAC can be seen as in Figure 3.6.

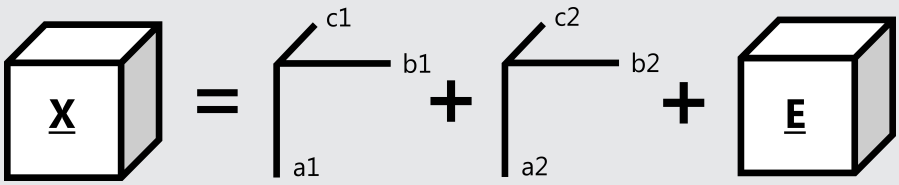

Figure 3.6: Graphical illustration of a three-way PARAFAC model.

Whereas bilinear methods such as principal component analysis have rotational freedom leading to an infinite number of solution, PARAFAC does not, leading to one unique solution. Obtaining a unique solution leads to unambiguous decomposition which obviously is a very attractive feature.

However, to reach a valid PARAFAC solution three requirements must be met: 1, data must be trilinear 2, the noise must be random and not too severe 3, the chemical rank of the system — the number of independently varying chemical components — must be estimated correctly. Typically excitation emission fluorescence spectroscopy is a well behaving trilinear system where the underlying emission and excitation spectra of the chemical phenomena can be retrieved (see Box 5 on page 61); GC-MS is also trilinear under ideal conditions (see e.g. **Paper IV** for more information).

In Figure 3.8 the temperature induced variation can be seen. The data cube had a chemical rank of three (water, urea and lactose) and thus a three component PARAFAC model was fitted. The scores and loadings of the model can be seen in Figure 3.9. From the spectral loadings the first component (blue line) is identified as urea and the corresponding scores accurately describe the urea concentration as seen in Figure 3.10. The remaining two components in the spectral loadings are very similar from 2070 – 2225 nm after which lactose shows its specific signals around 2270 and 2294 nm. Inspecting the temperature mode urea seems to be temperature independent, in other words, the urea signal does not shift over the

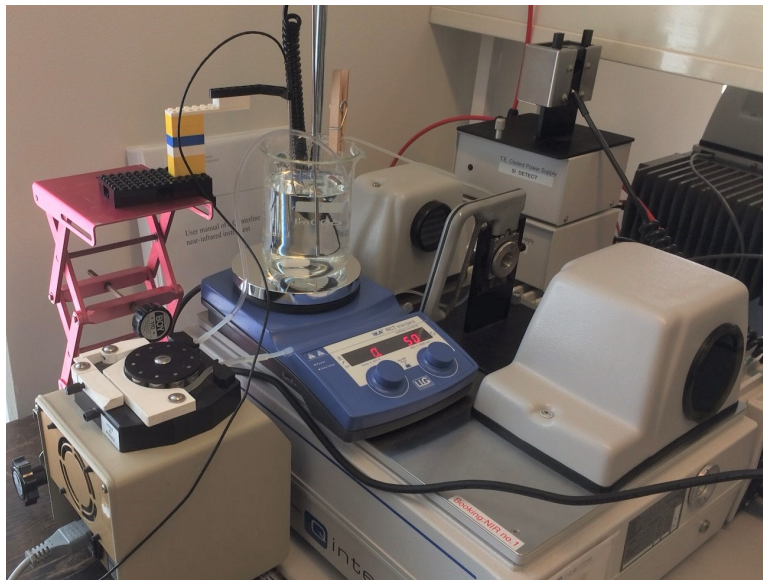

Figure 3.7: The experimental set-up used to acquire the 3-way NIR dataset.

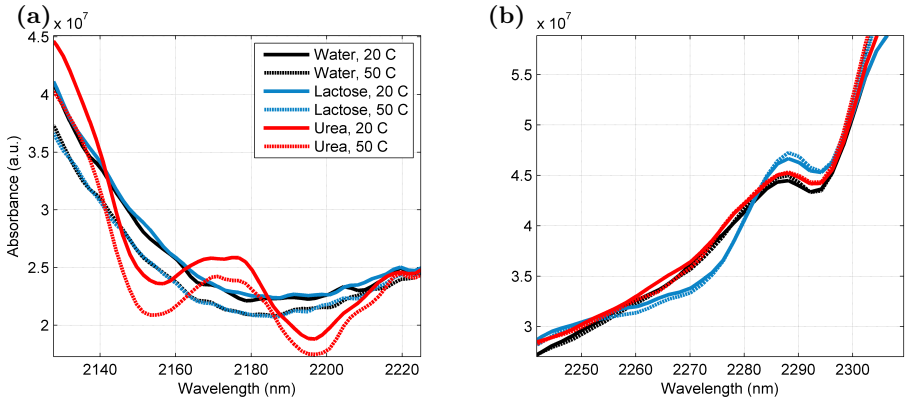

Figure 3.8: Second derivative near infrared spectra of water and solutions of water with lactose and / or urea. Urea absorption bands at 2155 nm and 2197 nm (a). Lactose absorption band at 2270 nm and an absorption band which was present in all samples at 2294 nm (b).

Table 3.1: The exact concentrations of the laboratory samples used to generate the 3-way NIR dataset.

Sample No.	Urea conc. (ppm)	Lactose conc. (ppm)
1	5252	956
2	2221	4172
3	3148	0
4	1037	0
5	0	0
6	0	5445
7	0	2063

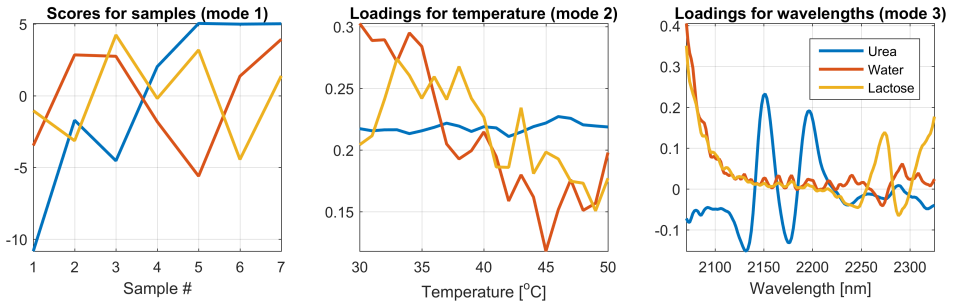

Figure 3.9: Three component PARFAC model with the seven samples at 21 different temperatures from 2070 - 2325 nm. Core Consistency 51.

temperature range used. Water and lactose NIR signals, however, are temperature dependent and spectrally overlapping. Different attempts to separate the water and lactose signals were performed using a PARAFAC model with constraints in the temperature mode. The idea was to force lactose and water to have different temperature loadings ideally leading to deconvolution of their spectral loadings. One PARAFAC component in the temperature mode was constrained to be linear and increase over the temperature development, which is the opposite of the water temperature dependency seen in Figure 3.9. However, this broke the correlation between the urea concentration and any of the PARAFAC components (results not shown). From a spectroscopic point of view, lactose and water may interact (by e.g. forming hydrogen bonds), meaning that it may be impossible to separate lactose and water. This is also hinted at by a low core consistency* of 51.

*The core consistency is a measure of how appropriate the PARAFAC model is. A core consistency of 100 indicates perfect fit [79]

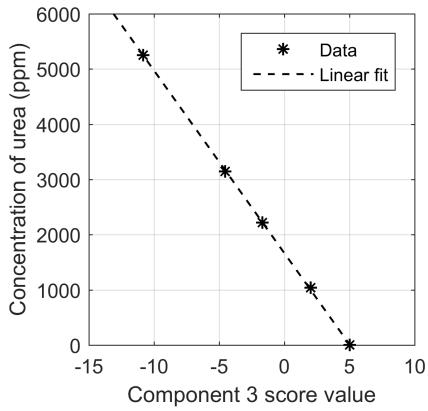

Figure 3.10: Correlation between concentration of urea and scores from three component PARAFAC model.

In conclusion the three component PARAFAC model showed the urea signal to be unaffected by temperature and this information could be efficiently extracted from the data cube. However, the lactose signal was found to be so strongly interconnected with the water signal that they could not be estimated independently by PARAFAC.

The spectral range chosen for the detection of urea has been based on work done in clinical research. Another spectral range may have been identified by contrasting high concentration urea sample to water and finding the NIR active areas. It can be expected that both the shoulder of the second overtone of water which carries information about N-H and the short-wave NIR range (950-1050 nm) also carrying information about N-H could be used instead.

It should be noted that no comparison with PLS prediction performance has been made since the concentration range in this experiment is way above the concentrations investigated in the **Paper II**. The purpose of this experiment was to explore the feasibility of using three-way NIRS method to overcome the temperature challenge. It should also be noted that setting up a commercial / professional three-way NIRS solution modulating temperature may turn out to be too complex for daily practise.

Chapter 4

Characterising evaporator condensate

In the production process of protein and lactose ingredients, UF or RO retentate are concentrated further via tubular falling film evaporators (Figure 4.1). In this type of evaporator system the feed enters at the top and is spread over a calandria (or *liquor distributor*) which is steam heated on the outside. The feed is pre-heated to the operating temperature and is introduced into the unit where evaporation takes place under vacuum, thus lowering the evaporation temperature resulting in gentle processing. The calandria spreads the feed in thin films to increase the surface to volume ratio thereby increasing the evaporation rate. Evaporator plants are usually built as several units in series concentrating the product more and more as it travels along the system, maximizing energy efficiency by heat recovery over the successive stages. The vapour is collected as condensate and the concentrated feed is sent to a spray drying operation. This evaporator condensate is another process-water recovery source and will be the topic of this chapter.

Three evaporator plants are running at AFI: Lactose2 (L2) is the newest and largest plant, while Wiegand1 (W1) and Wiegand2 (W2) are older and smaller. W1 and W2 each have two sampling points that make it possible to collect condensate from the first evaporator stage, where the product is not as concentrated as further downstream, and from the second (and last) evaporator stage. Condensate streams are combined up-stream from the sampling point for L2 resulting in samples that are composites of condensate from all evaporator stages in this plant.

Similar to the membrane permeate in **Paper I**, the evaporator condensate was analysed with the less biased and un-targeted GC-MS and ICP-OES approach. Several spectroscopic methods were also tried in an attempt to capture the quality of the condensate over the course of full production runs on industrial evaporator plants spread over two investigations.

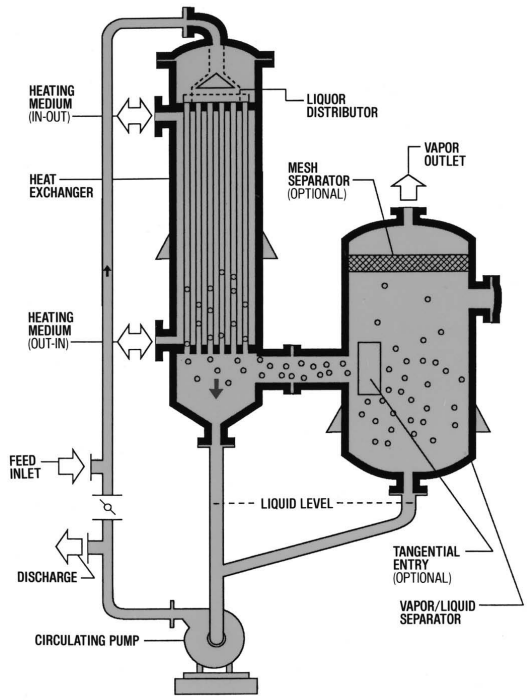

Figure 4.1: Schematic example of a circulating falling film evaporator from Buflovak.[80]

4.1 First investigation

4.1.1 Materials and methods

The first investigation on evaporator condensate was focused around the L2 plant. Twenty-four samples were collected in 250 ml sterile, amber, polypropylene sample bottles (ISOLAB, DE) over the course of a full production run (20 hours). Conductivity (HI2030-01 edge with conductivity probe, Hanna instruments, US), COD (2 ml water sample analysed as is with low range test tubes (0-150 mg $O_2 \cdot l^{-1}$) digested in a AL125 thermoreactor at 150°C for 120 minutes and measured in a AL200 COD VARIO photometer all from Aqualytic, DE), NIR spectra (same setup as in **Paper II**), ICP-OES (same setup as in **Paper I**), fluorescence EEM spectra (F900 Edinburgh Instruments UK, excitation: 250–450 nm in 5 nm steps, emission: 252–500 nm in 2 nm steps, 0.1 sec integration time, 1st and 2nd order Reyleigh scatter removed) and UV-Vis spectra (Evolution 220 UV-visible spectrometer from ThermoFischer Scientific, quartz cuvette with 1 cm path-length, 190-1100 nm in 2 nm steps, 0.4 sec integration time) were measured on all samples while infrared spectra (ABB Bomem MB 100 FT-IR (CA) with attenuated total reflectance crystal, 8 cm^{-1} resolution, 64 scans), ammonia / urea and lactose

(both kits from MegaZyme, IRE with 0.20 and 2.00 ml sample volume, respectively) were measured on a selection of samples.

Box 5: Ultraviolet-visible and fluorescence spectroscopy

Ultraviolet-visible (UV-vis) and fluorescence spectroscopy are described together in this box as they can be considered complementary techniques. A detailed account has been done by Atkins and de Paula.[81]

Similar to NIRS (Box 3) UV-vis spectroscopy concerns excitation of molecules and absorption of light. In UV-vis spectroscopy electrons are excited from the *highest occupied* to the *lowest unoccupied* molecular orbital. This requires far more energy than inducing fundamental vibrations and thus photons absorbed are in the visible or UV range. UV-vis active compounds are called chromophores and most often contain carbon-carbon double bonds, carbonyl or aromatic rings.

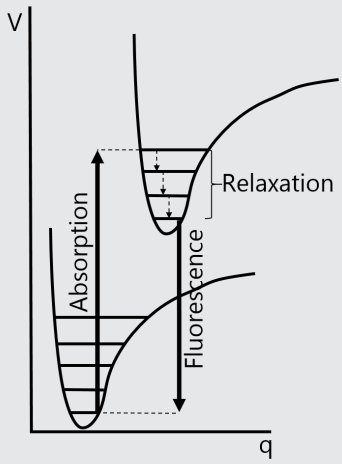

Figure 4.2: Jablonski diagram of electron excitation and potential photon emission.

Ordinarily the energy absorbed by a chromophore is degraded into thermal motion. However, some molecules will emit a photon instead — these molecules are known as fluorophores. Figure 4.2 illustrates the energy diagram of UV-vis absorption and the fluorescence emission of a photon.
In UV-vis spectroscopy the transmittance is measured while in fluorescence spectroscopy the emitted light is measured perpendicular to the incident light. A sample will be excited with a number of wavelengths and an emission spectrum for each can be collected into fluorescence landscapes also known as excitation emission (EEM) landscapes. In this configuration fluorescence spectroscopy focuses on the light emitted from the sample, but will also collect the scattered excitation light known as Rayleigh scattering as well as Raman scatter.

4.1.2 Results and discussion

The element concentrations in the evaporator samples were below the detection limit (dl, reported for all elements in parenthesis in this paragraph) for aluminum (13.6 µg·l^{-1}), bismuth (34.3 µg·l^{-1}), calcium (5.5 µg·l^{-1}), cadmium (20.3 µg·l^{-1}), cobalt (5.8 µg·l^{-1}), chromium (6.2 µg·l^{-1}), copper (13.6 µg·l^{-1}), magnesium (24.6 µg·l^{-1}), manganese (11.1 µg·l^{-1}), nickel (1.5 mg·l^{-1}) and strontium (1.57 µg·l^{-1}).

Only sodium (35.0 µg·l^{-1}), potassium (4.6 µg·l^{-1}) and lithium (70.6 µg·l^{-1}) could be quantified in the evaporator samples; the time series are shown in Figure 4.3.

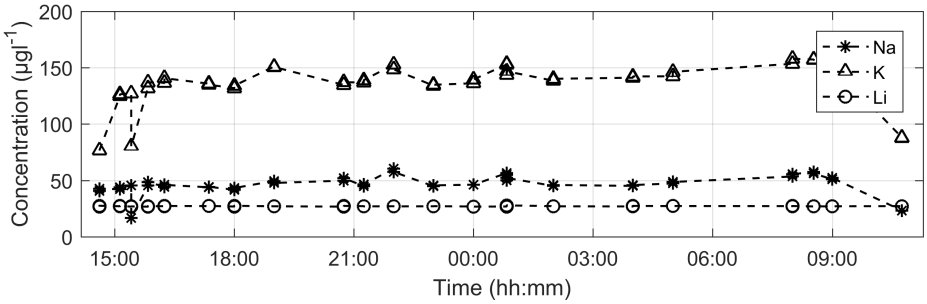

Figure 4.3: ICP-OES estimated concentrations of sodium, potassium and lithium in evaporator condensate samples from L2 collected in the first investigation over the course of a full production run.

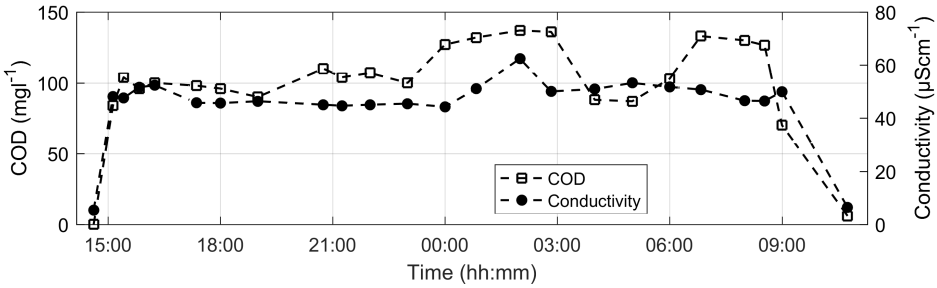

Figure 4.4: COD and conductivity measurements on evaporator condensate samples from L2 collected in the first investigation over the course of a full production run.

Lactose quantification was performed for six samples and all were below the detection limit (0.7 ppm). The same six samples were analysed for ammonia / urea. While urea concentrations were below detection limit (1.3 ppm) for all samples the ammonia concentration was estimated to be between six and eight ppm (detection limit 0.7 ppm). Infrared and NIR spectra were not able to show any meaningful results likely due to very low concentrations of compounds other than water, and are left out of further reporting.

COD and conductivity in Figure 4.4 showed approximately the same pattern where values increase after start-up to continue stable. The fluorescence signals (see Box 5) were dominated by the Raman scatter of water and to remove this artefact an EEM landscape of Milli-Q water was recorded and subtracted from each sample EEM landscape. This correction was necessary to obtain useful PARAFAC models but also had the adverse effect that shot-noise increased as can be seen

in Figure 4.5d. The EEM landscapes were decomposed with a two component PARAFAC model (see Box 4). Component 1 shows a dynamic behaviour similar to the COD with a small upwards shift after 12 hours of production. Identification of the two fluorophores is challenging as they do not resemble known individual fluorophores, but rather complex signals as explained by e.g. Henderson et al.[56]

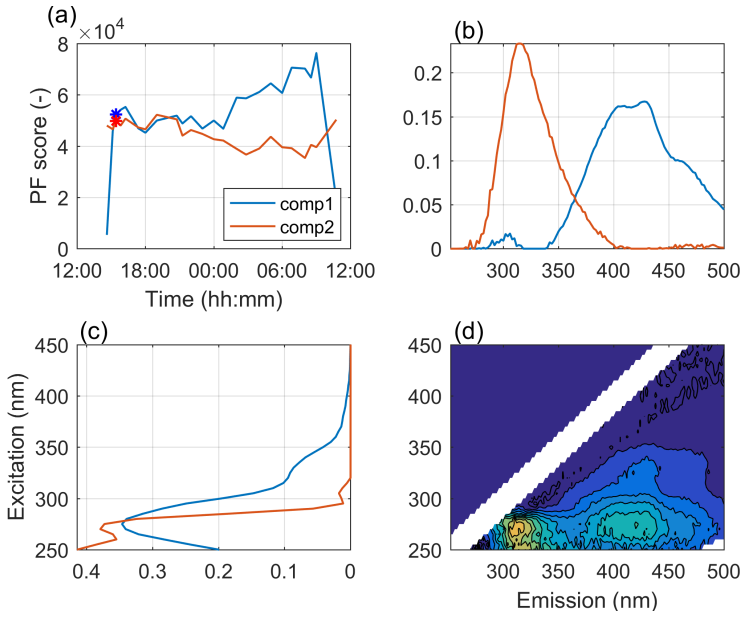

Figure 4.5: PARAFAC decomposition of fluorescence EEM spectra obtained from evaporator condensate samples from L2 collected in the first investigation over the course of a full production run. PARAFAC scores for the two components (a), emission loadings for the two PARAFAC components (b), excitation loadings for the two PARAFAC components (c), and an example of a EEM landscape of the sample taken at the time-point marked by an asterisks in top left plot (d).

The obtained UV spectra (see Box 5) only showed absorbance around 240–300nm as can be seen in Figure 4.6a. The peak maxima over the production run after an off-set correction using absorbance at 975 nm can be seen on the right. A low signal to noise ratio makes it hard to evaluate if the dynamics are similar to the dynamics seen in the COD and fluorescence data.

It seems that the evaporator condensate contains some organic compounds that fluoresce and absorb light around 260–280 nm in the UV region. Chemically this could indicate that aromatic amino acids were present in the condensate samples. For fluorescence the signal could not be enhanced significantly without increasing the acquisition time, and since this was already high (approximately 30 minutes

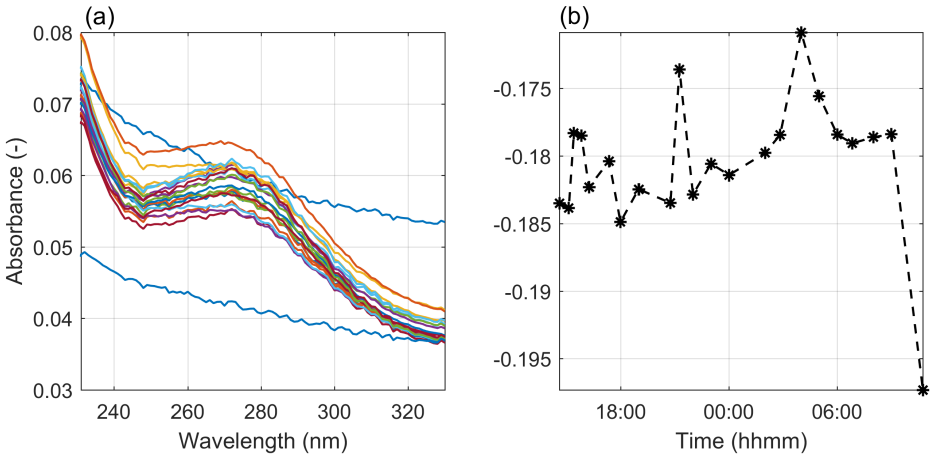

Figure 4.6: UV spectra obtained from evaporator condensate samples from L2 collected in the first investigation over the course of a full production run (a), absorbance at 272 nm after an off-set correction at 975 nm (b).

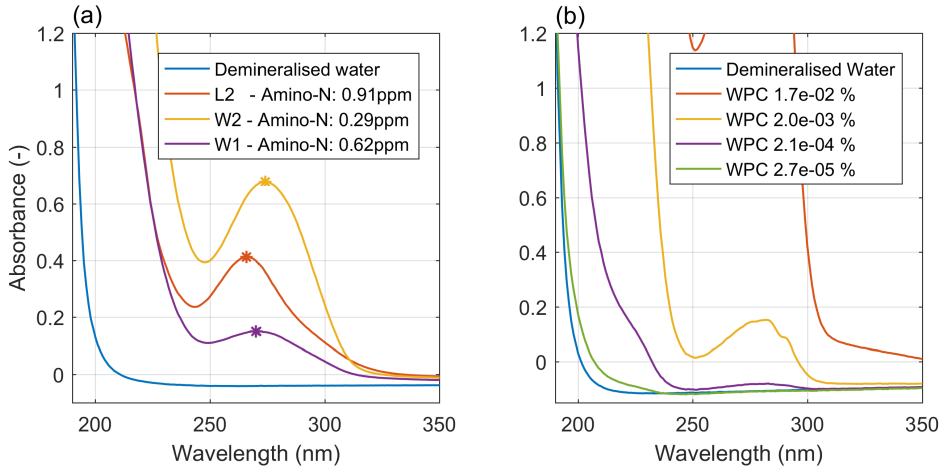

Figure 4.7: UV spectra with 10 cm path-length of three evaporator condensate grab samples obtained from three evaporation plants at AFI (a). In the legend the amino-N quantification are added. The maxima between 250–300 nm of the three samples are marked with asterisk (*). UV spectra with 10 cm path-length of a whey protein concentrate dilution series (b).

per sample) another approach was sought. UV spectroscopy also detected a signal and to increase sensitivity a cuvette with 10 cm path-length was tested. Three new evaporator condensate grab samples — one from each evaporator plant — were collected and tested on the new UV setup and the results are shown in Figure 4.7a.

An amino-N enzyme kit (Megazyme, IRE; quantifies the terminal nitrogen of amino acids and peptides) was also tested as amino acids were suspected to give rise to the signal. The hope was that the UV signal would be able to predict amino-N concentration in the condensate. Based on Figure 4.7a there is no obvious relationship between the amino-N quantifications and the intensity of the aromatic amino acids peak around 270–280 nm across the three processing plants. The issue can be analogous to what Henderson et al.[56] described in the relationship between fluorescence and TOC in that not all amino acids give rise to a UV signal, and if the ratio between UV-inactive and UV-active amino acids changes over time, from plant to plant or from feed to feed, it cannot be related directly to the UV signal. The change in amino acids composition is also hinted at in the peak shift from evaporator plant to evaporator plant. The amino-N quantifications are also affected by changes in length of proteins and peptides, which could disturb the relationship.

Unfortunately the amino-N quantifications came with a high uncertainty. From the results shown in Figure 4.7a each sample was measured in triplicates. The standard deviations were pooled and the 95 % confidence limit was estimated to be ± 0.11 ppm for the method.

To get an idea of the amino acid concentration / protein level in the evaporator condensate a dilution series of whey protein concentrate (80 % whey protein, WPC80) was measured on the UV spectrometer (Figure 4.7b). The dilution that resembles the evaporator condensate samples most is the 0.002% WPC sample (orange line). This indicates that the concentration of amino acids is around 16 ppm WPC equivalents ($= 0.8 \cdot 0.002\%\text{WPC} = 0.8 \cdot 0.02 g\text{WPC} \cdot l^{-1} = 0.8 \cdot 20 ppm$ WPC)

In order to identify the fluorescing and UV active compounds GC-MS was performed on the three grab samples with the same method as described in **Paper I**, with the exception that the sample workup was modified. Starting sample volume was 10 ml in new 50 ml falcon tubes, sample dried in freeze-drier, re-suspended in 1 ml Mili-Q water, 100µl transferred to GC inserts and dried again before derivatization. The identified (to level two as in **Paper I**[82]) compounds are reported in Table 4.1.

Table 4.1: Identified compounds from first and second investigation. Symbols: - indicate signals found to be less than the LOD (calculated as the mean of the blanks plus five times the standard deviation of the blanks) for all samples; + indicates signals found larger than LOD in at least one measurement from a given sampling point.

		1st trial			2nd trial		
Compound	**Rt**	**L2**	**W1[b]**	**W2[b]**	**L2**	**W2[a]**	**W2[b]**
Butanoic acid	4.80	-	-	-	-	-	+
Di-methyl ethanamine	4.99	-	-	-	-	-	+
Butanediol	6.50	-	-	+	-	-	+
Propanediol	6.59	-	-	+	-	-	-
Lactic acid	6.76	+	+	+	+	-	+
Pyruvic acid	7.01	-	-	-	+	-	-
Hexanoic acid	6.89	-	-	+	-	-	-
Alanine	7.15	-	-	+	-	-	-
Glycine	7.33	-	-	+	-	-	-
β-hydroxybutyric acid	7.69	+	+	+	-	-	-
3-methyl butanoic acid	7.83	-	-	+	-	-	-
Hydroxyisocaproic acid	8.38	+	+	+	-	-	-
Benzoic acid	8.57	-	+	+	-	-	-
Urea	8.64	-	-	+	-	-	-
Octanoic acid	8.72	-	+	+	-	-	-
Phosphoric acid	8.79	-	-	+	+	-	+
Benzene-diol	9.14	-	-	-	+	-	-
Succinic acid	9.14	-	-	+	+	-	-
Uracil	9.43	-	-	-	+	-	-
Glyceric acid	9.27	-	-	+	-	-	-
Decanoic acid	10.44	-	+	+	-	-	-
α-hydroxy glutaric acid	11.38	-	-	+	-	-	-
Pyrimidinecarboxylic acid	12.69	-	-	-	+	-	-
Phosphoric acid propandiol	12.99	-	-	-	-	-	+
Citric acid	13.51	-	-	-	-	-	+
Sugar	13.51	-	+	+	+	-	+
Lactose	18.25	-	-	+	-	-	+

[a] Sampling the first stage of the evaporators.
[b] Sampling the second stage of the evaporators.

4.2 Second investigation

Based on the results from the first investigation of evaporator condensate above, a second investigation was performed focussing on UV spectroscopy, amino-N quantification and GC-MS.

4.2.1 Materials and methods

Sixty-one samples were collected from 13:35 till 13:37h the following day, again covering one full production run of approximately 24 hours. UV spectra (path-length 10 cm) and GC-MS was measured for all samples with the methods described previously and amino-N was determined for twelve samples as described above. Furthermore, two grab samples were taken from W2 from both the first and second stage.

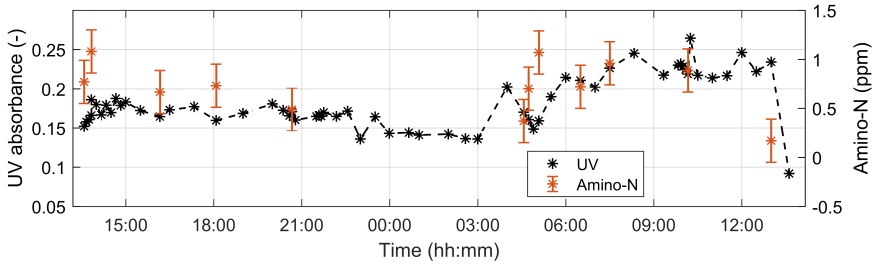

Figure 4.8: Maximum UV absorbance between 250–350 nm obtained from evaporator condensate samples from L2 collected for the second investigation over the course of a full production run. Amino-N quantifications with 95 % confidence bands.

4.2.2 Results and discussion

The UV absorbance and amino-N quantifications are presented in Figure 4.8. Due to the large uncertainty of the amino-N analyses and the limited process variation over time it is hard to determine whether the UV absorbance correlates with the amino-N quantifications.

Interestingly, the UV absorbance makes a small upwards step around 06:00h similar to what was seen in COD and fluorescence in the first investigation. Figure 4.9a shows the two signals superimposed and on a relative time-from-start-up axis rather than calendar time. To emphasize the similarity in dynamic behaviour the time-axis is converted to per-centage of the complete production run in Figure 4.9b. It seems that UV and fluorescence are capturing some time dynamic phenomena related to the L2 plant from two independent runs. It is reasonable to expect that UV and fluorescence are capturing the same underlying phenomena related

to aromatic amino acids as these compounds are active in both spectroscopic techniques.

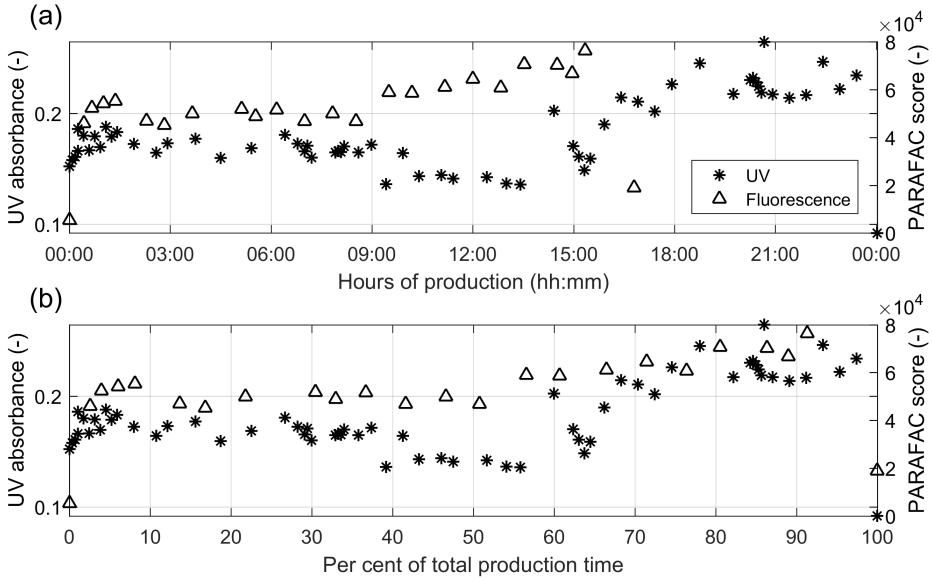

Figure 4.9: UV absorbance from second investigation and PARAFAC scores for second component from first investigation superimposed across their respective production run times (a). As above but plotted against per cent of total processing time (b).

No amino acids or peptides were detected with the GC-MS as shown in Table 4.1. As in **Paper I** the compounds are only reported as detected or not detected relative to a detection limit. The samples were prepared as six replicates where some of the chromatograms for the six replicates occasionally provided meaningless data, possibly due to moisture migrating into the sample vials as a result of incomplete sealing. Moisture reacts with the derivatization agent consuming it before other compounds can. Two examples of why only detect / non-detect is used for reporting are shown in Figure 4.10. Lactic acid is quantified frequently above the detection limit, while pyruvic acid is only occasionally observed. The replicates for a given sample are sorted according to time of analysis, which can be seen clearly for e.g. sample *1015* with lactic acid. Here a clear decline in signal as function of time of analysis is apparent. For most samples the first replicate gives the largest signal, underlining that the compound's concentrations are close to the detection limit.

The evaporator W2 was sampled from both the first and final stage. In the first stage no compounds could be detected in the samples, while some compounds were found in samples from the second stage (Table 4.1). This suggests that the

evaporator condensate quality deteriorates as the product stream becomes more concentrated.

From Table 4.1 it is clear that lactic acid is the most often found compound. Vourch et al.[31] analysed five condensate samples from different processes and quantified lactic acid together with lactose, ethanol, acetone and acetoin. The concentrations reported are all in the $\mu g \cdot l^{-1}$ range. These compounds are very volatile (except lactose) meaning that acetone, ethanol and acetoin will not be detected with our presented method. Acetone and ethanol accounted for far the most organic material of the measured compounds in the analysed evaporator condensates by Vourch et al.[31]. This suggests that the work-up procedure for our presented GC-MS method is not ideal as many volatile compounds will be overlooked.

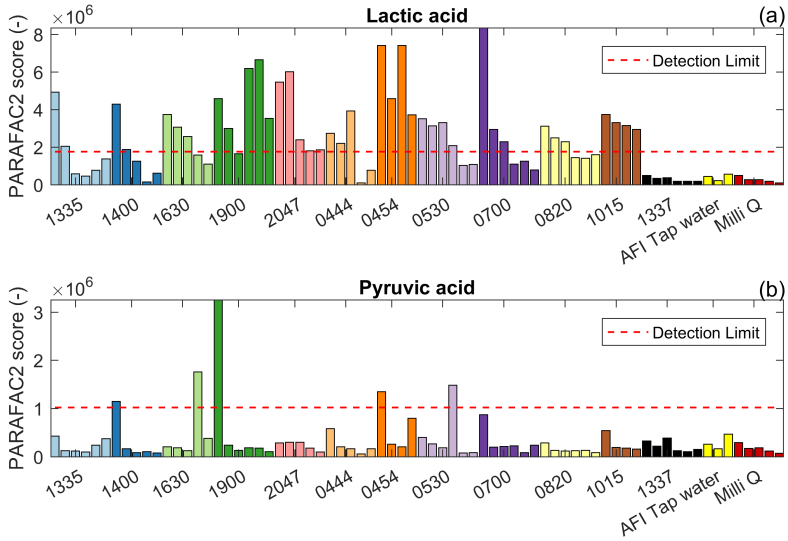

Figure 4.10: PARAFAC2 scores for selected compounds for all replicates of the eleven evaporator condensate samples analysed with GC-MS. Replicates order according to analysis sequence. Samples collected during the second investigation. Lactic acid (a) and Pyruvic acid (b).

Recently solid-phase microextraction (SPME) technology has emerged as an alternative and automated sampling / work-up method in analytical chemistry.[83] Instead of freeze-drying the samples to concentrate compounds, a polymer coated object is emerged into the sample or presented to the headspace of the sample. The compounds will then migrate into the polymer, effectively up-concentrating the compounds; later they are released directly into a GC with a thermal desorption unit. Stir-bar (magnetic stirrer bar coated in the SPME material, usually

polydimethylsiloxane also known as PDMS), also know as stir-bar sorptive extraction (SBSE), has been used for quantification in wastewater[83] and food[84] samples with impressive detection limits in the *ppb* range. The reason for the high performance is the fact the much more polymer can be used to coat the stir bar compared to for example a single fibre typically used in aroma analysis. However, in order to truly be quantitative matrix effects must be handled by e.g. matrix matching calibrations, use of internal standards, and / or standard addition. Hopefully, this is less challenging in process-water since the matrix is modest in complexity compared to e.g. food and wastewater. To establish a SBSE method some work will be required to find optimal extraction conditions (salt concentrations, temperature, time, pH and more).[85]

4.2.3 Conclusions

So far it has not been possible to determine a key quality attribute for evaporator condensates from AFI. UV and fluorescence spectroscopy indicate that amino acids or peptides are present in the water streams, but quantification of these compounds has not yet been successful.

In future studies alternative work-up strategies should be considered focusing on quantifying volatile organic compounds and amino acids to avoid overlooking potential quality attribute candidates as the current methods may have done. Stir-bar sorptive extraction seems to be an obvious method due to the (reported) high sensitivity.

The evaporation processes investigated here appear, similar to RO membrane filtration processes, to be quite stable over production runs. However, this is under the assumption that fluorescence and UV spectroscopy do capture some important quality aspect of the process-water.

Chapter 5

Extreme observations in food processing

In **Paper I** urea was identified as the main organic compound permeating the membranes. **Paper I** and **Paper II** suggested that the investigated processing plant performed very consistent, hence was running under *Normal Operating Conditions* (NOC). However, the additional data presented in Chapter 2 (Figure 2.6 on page 29) documenting two occasions of lactose permeating through in quantifiable concentrations might suggest that the process may be disturbed, or was running under *Abnormal Operating Conditions* (AOC) from time to time. All processes will occasionally run under AOC due to (external) disturbances and for this processing plant it appears that the lactose concentration occasionally increases in the permeate.

In **Paper I** the growth potential was investigated on actual process-water samples under NOC. However, it could be relevant to perform these experiments on process-water samples collected during AOC or mimicking AOC, leading to the best imaginable growth conditions for microbes or a worst-case scenario in terms of process-water storage time. If a worst-case scenario is to be defined for further microbiological work and risk assessment, sampling campaigns may have to run for extensive amounts of time to — by chance — find such a (worst-case) situation. The worst-case scenario in this case is defined by the highest concentration of urea and lactose that can be expected (within a defined time-frame).* Fortunately, statistical tools have been developed to characterise and estimate such worst-case scenarios and these will be the main topic of this chapter.

5.1 Extreme value theory

In most measurement based sciences extreme observations are habitually considered outliers as they do not follow the bulk of the data acquired. Due to the infrequent nature of extreme observations they are often disregarded in statistical

*Since urea is retained poorly disturbances will not affect the concentration significantly and lactose should probably be the focus.

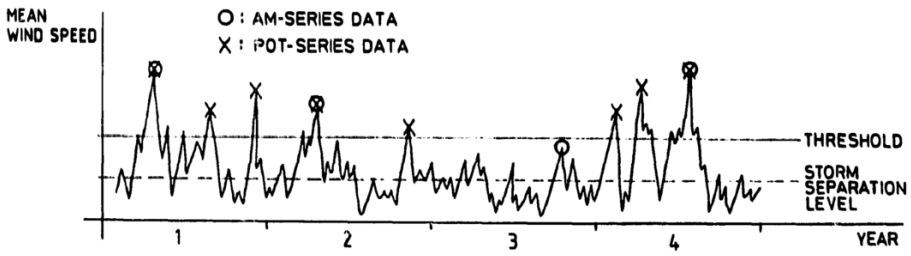

Figure 5.1: Illustration of the block maxima (here denoted Annual Maxima, AM) and point over threshold (POT) method by Abild et al.[88]

evaluations. In some scientific fields — e.g. hydrology — focus has naturally been on the extreme situations. Extreme rainfall and flooding both have major impact on those affected by it, and describing these situations has been of much interest.[86] Extreme value theory has evolved from mathematical statistics and Gumbel was among the first to present statistical applications.[86,87]

In EVT the first step is to separate the extreme observations from the non-extreme observations. For EVT to give reliable estimates these extreme observations must be independent and identically distributed (IID). Generally speaking there are two approaches in extracting extreme observations: 1, *block maxima* method (BM) and 2, the *point over threshold* method (POT). Both methods were illustrated by Abild et al.[88] reproduced in Figure 5.1. Next some aspects of EVT will be explained and illustrated through laboratory generated data as well as simulations.

Laboratory experiment In order to test out EVT a laboratory experiment was set up in which liquid was drawn from two separate vessels by separate pumps. One containing 250 ml juice (Ribena), 25 ml 5M HCl and 8l de-mineralised water, and one containing a 0.9 % NaCl solution. The two liquids were joined in a T-junction and mixed before measuring pH. The pumps drawing the solutions were turned on and off in random patterns, switching every 20 seconds to induce random variation, using a USB control unit (K8055 from Velleman inc., BE). The pH value was logged every fifth second to generate the data set seen in Figure 5.2.

We could think of this laboratory experiment as a process signal of a pH adjustment step in e.g. a continuous membrane separation process. It may be paramount that pH is kept below a upper-desired value and it is important to know how often — if ever — pH rises above e.g. 3 during a given time-span, say 600 minutes (pH is recorded every 5 seconds so this is equivalent to 7.200 observations). Control charts can be put in place to warn operators to take action as shown in Figure 5.3a, but this does not give a complete description of the frequency with which the situation will potentially occur.

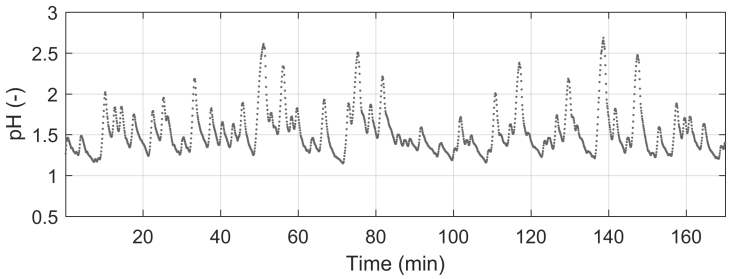

Figure 5.2: pH values obtained from laboratory experiment drawing two liquids at random and mixing.

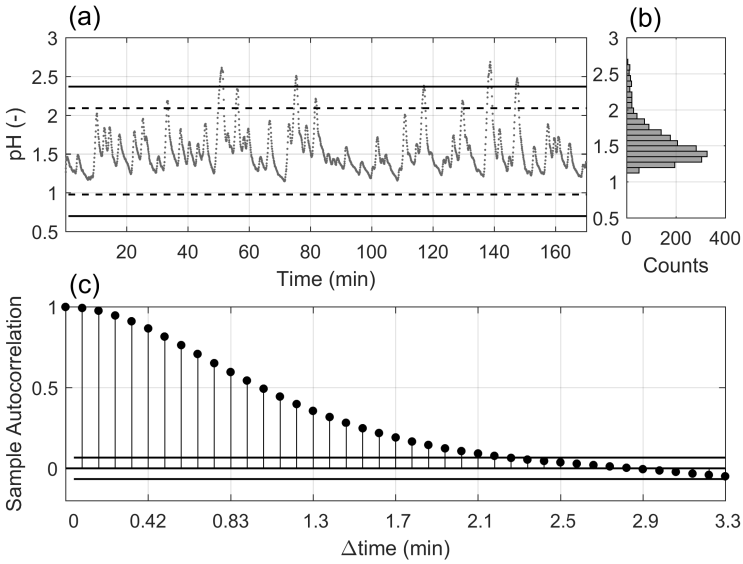

Figure 5.3: Control limits (two and three times the standard deviations) added to the pH values from the laboratory experiment (a), as well as a histogram of the data (b) and the autocorrelation of the pH signal (c).

In the laboratory experiment EVT was applied using the BM approach blocking 26 (= 2.2 minutes) consecutive observations together. This led to overall 79 extreme observations which are modelled with the generalised extreme value (GEV; explained in **Paper III**) distribution shown in Figure 5.4a. The pH (return-level) that on average can be expected during 600 minutes of processing is

$$T = 7200/26 = 277$$
$$3.4 = F^{-1}(1 - T^{-1})$$

Notice that 7200 observations represents the time, T, of 600 minutes, but must be in the same unit or scale as the extracted maxima. Since an extreme was

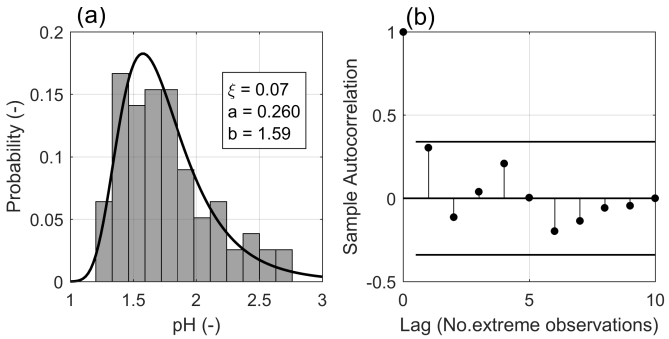

Figure 5.4: Generalised extreme value distribution fitted to block maxima extracted extreme values from the laboratory experiment (a) and the autocorrelation between the extracted extremes (b).

extracted every 26^{th} observation, the time must also be converted back. The average time (return-time) between observing pH values of 2.5 can be estimated via the cumulative distribution function as

$$T = (1 - F(x))^{-1} = (1 - F(2.5))^{-1} = 22$$
$$48min = \frac{22blocks \cdot 26obs/block \cdot 5sec/obs}{60sec/min}$$

Again this time estimate is in unit *block* as for the return-level. This is converted to observation by multiplying with the block size to get the estimate in terms of observations, and since observations are recorded in 5 seconds intervals it must be multiplied with five and then divided by 60 to obtain the estimate in process time.

Extreme value theorem The extreme value theorem (also known as the Fisher-Tippett-Gnedenko or Fisher-Tippett theorem) states that regardless of the underlying distribution of the original data set, the (IID) extreme observations will follow three and only three extreme value distributions: Gumbel, Fréchet or Weibull.[87,89] Fisher and Tippett gave the key results in 1928, while Gnedenko published the formal proof in 1943.[87]

The fact that extremes extracted from any distribution will lead to one of the three extreme value distributions makes it the extreme value analog to the central limits theorem.[89] Figure 5.5, 5.6 and 5.7 show examples of the theorem in action by sampling extremes from three very different parent distributions (triangle, normal and log-normal) and using the BM method leads to very good fits for the GEV distribution. Inferring which of the three extreme value distributions the extreme samples follows must be done via the tail-parameter, ξ.

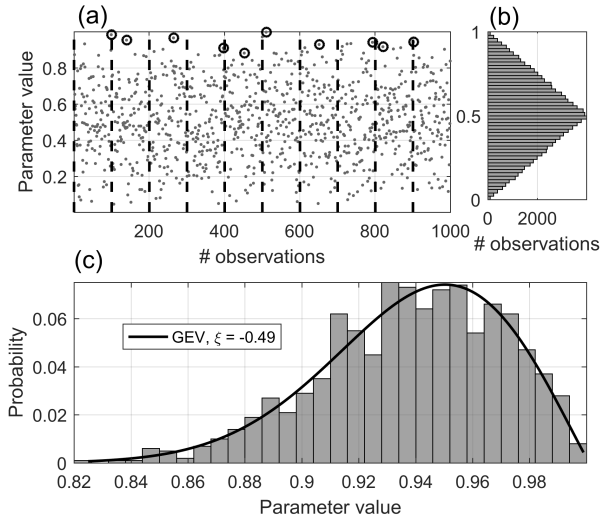

Figure 5.5: Example of the extreme value theorem. 100.000 observations drawn from a triangle distribution (b), blocking data into subsets of 100 observations and extracting maxima from each block (a) and the resulting extreme value data set and generalised extreme value probability distribution fitted onto (c).

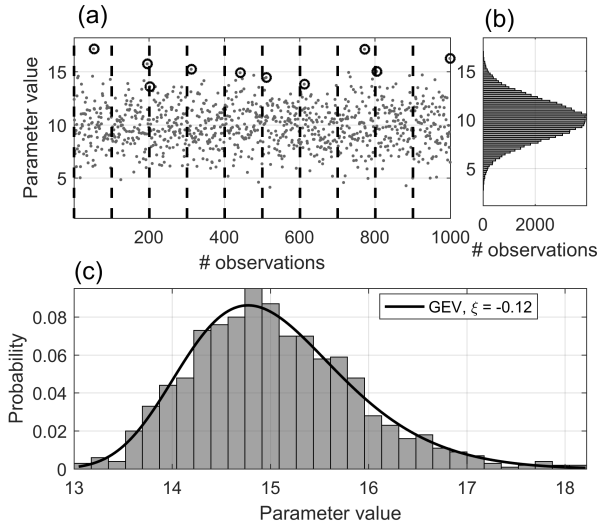

Figure 5.6: Example of the extreme value theorem. 100.000 observations drawn from a normal distribution (b), blocking data into subsets of 100 observations and extracting maxima from each block (a) and the resulting extreme value data set and generalised extreme value probability distribution fitted onto (c).

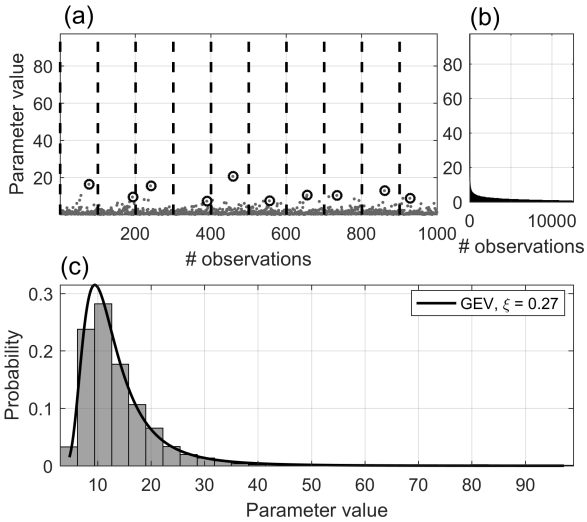

Figure 5.7: Example of the extreme value theorem. 100.000 observations drawn from a log-normal distribution (b), blocking data into subsets of 100 observations and extracting maxima from each block (a) and the resulting extreme value data set and generalised extreme value probability distribution fitted onto (c).

Some application of EVT EVT has been applied in many other scientific fields such as wind engineering for bridge building and estimating pipe corrosion:

- Bridge building:[88] To verify the structural integrity of the Great Belt bridge (DK) the fastest expected wind speeds across the Great Belt over the expected life-time of the bridge was estimated. To estimate the maximum wind speed a total of nine annual maxima were extracted from the complete data record of 9 years of consecutive 10-minute averages.
- Pipe corrosion:[90,91] Process pipes are subject to damage from corrosion, fatigue, creep and interaction between these three factors. To ensure that the pipe's integrity is not at risk, the thinnest / most damaged part of the pipe is estimated using EVT. Rather than having a full data record covering the entire pipe, the pipe will be inspected at different locations.

The two examples illustrate different scenarios in which to apply EVT. When studying pipe corrosion finding the thinnest part of the pipe is the goal and samples are taken to find this. Pipes are not sampled periodically in an evenly distributed or random fashion, but rather places that are expected to be more corroded are probably targeted, while at the same time being constrained by physical accessibility of the pipe (i.e. under water, under ground or otherwise inaccessible). In

contrast, wind speeds across the Great Belt were recorded every ten minutes year-round leading to a complete data record. In this view dividing the data into one year blocks (which was the accepted norm) could seem overly reductionist. Ten years before the paper by Abild et al., Cook[92] (inspired by Jensen and Franck[93]) suggested to extract the maximum from each *individual* storm rather than the annual maximum. The approach was named *Method of Independent Storms* (MIS). With MIS Cook identified on average one hundred storms per year, dramatically increasing the number of observations to use for the BM method.

Independence of observations From Figure 5.2 signal correlation is evident in the laboratory experiment and expected since the system will have some memory effect due to holding times and mixing. Figure 5.3 shows conventional control limits of two and three standard deviations (a), the histogram of the data set (b) and the autocorrelation (c).

Knowledge on the autocorrelation can help choose which EVT approach (BM or POT) is most appropriate since both assume IID extreme observations. POT is more naïve in this sense as it will over-estimate the number of extreme situations when autocorrelation is present simply due to the fact that extreme as well as non-extreme observations correlate. In the BM approach data is sub-sampled and only one observation per block is extracted. This leads to a smaller number of extreme observations, but also guarantees that they are independent (assuming the segmentation is done properly). In the pH process signal example the autocorrelation gives us a good indication that data should be sub-sampled in blocks of at least 26 observations (2.2 minutes), which lead to uncorrelated extreme observations. Cook[92] showed one way of doing this for wind speed and we show another example for food processing in **Paper III**.

Blocking has another advantage other than (ideally) ensuring independent extreme observations; it can also help *target* the analysis. Targeting should be understood in the sense that several underlying phenomena can be investigated. For example, in a production facility producing 24 hours a day, it may be interesting to investigate if all eight hour shifts have the same probability of extreme events and by stratified blocking this question can be addressed.

Identically distributed observations As mentioned before, the extracted extremes must also be *identically distributed*. This is an established assumption in regular statistical analysis. However, in EVT this assumption may be more difficult to satisfy seeing that usually many observations have to be collected over extended periods of time to ensure at least some extreme observations occur. Collecting data over extended time periods increases the risk of the parent distribution changing, e.g. climate change will effect precipitation patterns in future hydrology investigations.

EVT is very much a tool for risk assessment. Interestingly, it has seen very few (or no) applications within microbiological risk assessment based on an open literature search. Neither has EVT found its way (yet) into food processing technology. In **Paper II** we suggest how to apply EVT to both characterise food processing performance of a key quality attribute and microbiological load in the final product.

Paper III

Tutorial - applying extreme value theory to characterize food-processing systems

Peter B. Skou, Stephen E. Holroyd and Frans van den Berg

Journal of Chemometrics, 2017, e2896.

Received: 14 December 2016 | Revised: 8 March 2017 | Accepted: 9 March 2017

DOI: 10.1002/cem.2896

TUTORIAL

WILEY Journal of CHEMOMETRICS

Tutorial – applying extreme value theory to characterize food-processing systems

Peter B. Skou[1] | **Stephen E. Holroyd**[2] | **Frans van den Berg**[1]

[1] University of Copenhagen, Department of Food Science, Frederiksberg, Denmark

[2] Fonterra Research and Development Centre, Palmerston North, New Zealand

Correspondence
Peter B. Skou, University of Copenhagen, Department of Food Science, Rolighedsvej 30, Frederiksberg, Denmark.
Email: peter.b.skou@food.ku.dk

Funding information
Danish Council for Strategic Research; Commission on Health Food and Welfare

This tutorial presents extreme value theory (EVT) as an analytical tool in process characterization and shows its potential to describe production performance, eg, across different factories, via reliable estimates of the frequency and scale of extreme events. Two alternative EVT methods are discussed: point over threshold and block maxima. We illustrate the theoretical framework for EVT by process data from two different examples from the food-processing industry. Finally, we discuss limitations, decisions, and possibilities when applying EVT for process data.

KEYWORDS
Extreme value theory, food processing, return-level estimation, return-time estimation

1 | INTRODUCTION

Production without variation in areas such as food and dairy technology is a utopia. Despite the evermore sophisticated measurement and control strategies,[1] some residual uncertainty from both production technology and biological variation will always remain in the highly complex systems involved in food processing. During production, extreme values will occur in the process parameters periodically and thus have impact on final product composition. Even if no direct consumer risks are introduced under those extreme situations, a better description of and knowledge on these events can be a valuable asset from a management-level or control and optimization point of view. Despite extreme situations occurring infrequently by nature, the consequences can be significant of lost revenue, and with the ever-increasing scale of operations encountered in the food industry, the ability to characterize extremes is economically relevant. In process-monitoring extreme values of key-processing parameters are habitually considered outliers, or otherwise not regarded in a statistical sense. These very large deviations from the desired set-point (either on the low or high end) lie in the very tails of the statistical distribution. Normal and related distributions from classical statistics focus on the mean and the spread around the mean, and often do not describe tail behavior well.

Describing a process parameters' extreme behavior can be very informative: how often do we see a deviation from the set-point of this size (the so-called return-time), how large a deviation can we expect to see over the course of 1 production run (return-level), or are different production sites running the same process observing extremes at the same rate (plant- or company-wide optimization and alignment)? These are some of the answers and information statistical analysis of extreme observations can provide. By extreme value analysis (EVA) process, managers can acquire systematic and objective measures of the process performance that were previously unknown. Extreme value analysis is normally associated with risk management to perform probabilistic analysis of rare, but severe, events, in a much more informed way; here, we argue that EVA can be a relevant tool for everyday process management as well. Extreme value theory (EVT) contains the tool set to describe extreme observations. In many fields such as hydrology, wind engineering, material testing, and finance (see, eg, other studies[2-4]), these tools have been picked up and are used efficiently. However, there are very few publications dealing with the application of EVA/EVT principles and techniques to process characterization.

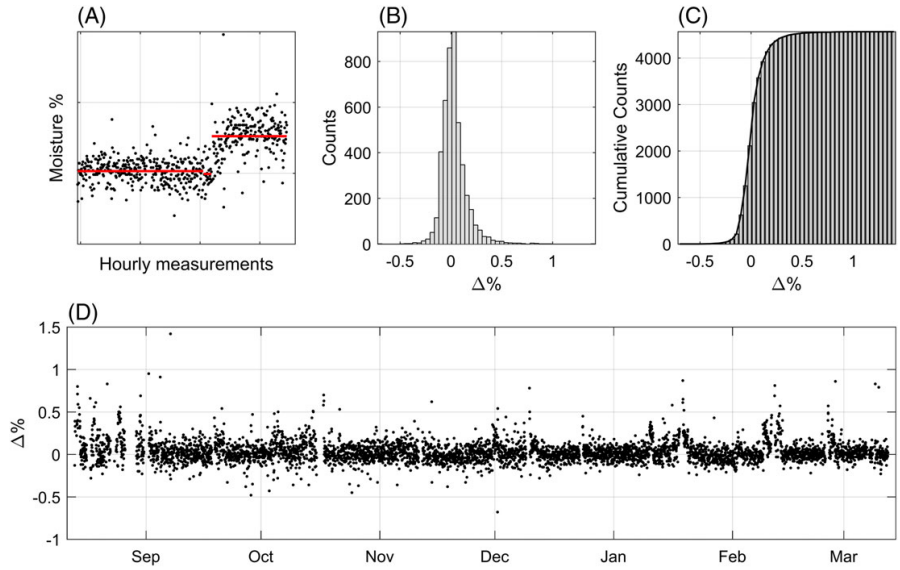

FIGURE 1 Factory 1a data (n = 4565): A, small extract from the data record with product dependent set-points (red line); B, histogram; C, empirical cumulative distribution function; and D, deviation from set-point over time

Moisture content is 1 key quality attribute in production of spray-dried milk powder. High moisture content leads to accelerated deterioration (and impaired shelf life), which may demand reprocessing of the batch or lowering the market quality grade, while low moisture content indicates drying more than was needed. Over drying thus implies spending extra energy and money on processing, and consequently selling more milk solids for the same price. It is therefore of utmost importance that the process parameter moisture content in milk powders is under control. While small variations from the desired moisture content may not impair powder quality nor impose considerable extra expenditure, large deviations (extreme observations) are likely to cost substantial amounts of money and should be visible (and to some extent predictable) from statistical inference. At Fonterra, gross composition, including moisture content, of whole milk powder is measured approximately every hour at-line by Near InfraRed Spectroscopy–based predictions.[5] For our investigation, moisture data from three separate production lines in two factories were collected from August 2015 to March 2016, and this data set will serve as an example on how to apply and interpret EVT in process characterization. In 1 facility, moisture data were collected from two parallel fluid beds (denoted Factory 1a and Factory 1b henceforth) that were fed from the same drier, while data collected from Factory 2 originate from a separate production line. To eliminate differences resulting from product specific target values, the data are transformed to percentage deviation from set-point*:

$$\Delta\% = \text{set}-\text{point}(\text{moisture}\%) - \text{measured}(\text{moisture}\%). \quad (1)$$

The data from Factory 1a are visualized in Figure 1. In this tutorial, we look explicitly at large positive values for $\Delta\%$, meaning powder dried more than is economically desirable.

The process parameter in Figure 1 can be considered a stationary series since the set-point is known for each observation from the processing recipe, and we can thereby remove—or detrend—the product-specific levels. If a set-point is not available, additional measures must be taken in the form of detrending or filtering.[6] This will however not be the focus of this tutorial. As first approximation, our hourly moisture measurements x—with numerous small and independent disturbances acting upon them such as sampling variation, analytical error of the NIR measurement, etc—could be thought of as following the standard normal distribution using the well-known Gaussian cumulative distribution function (CDF; Equation 2a) and probability density function (PDF; Equation 2b):

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx, \quad (2a)$$

*Note that the difference is chosen so that the extremes we want to investigate are positive. This is done for computational convenience; if the minima of the former are of interest the order could be switched.

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad (2b)$$

where μ is the distribution location (or expected mean) of x, and σ^2 represents the spread (or expected variance) in the data set. Note that there is no elementary function available for the CDF of a normal distribution, but we included an open form here to compare with other data distributions presented later on in the tutorial. The data from Factory 1a is fitted to 2a, and the results are shown in Figure 2.

Looking at the results in Figure 2, it is clear that the bulk of the data is modeled well by the normal distribution but that extremely high (and low) moisture deviations from set-point do not follow this expectation—a deduction that could already be made from the empirical distributions shown in Figure 1A,B and even the raw data in Figure 1C. The aim of this tutorial is to discuss modeling strategies and statistical inference on these tail extremes.

Extreme value theory and EVA are, despite their long history, only infrequently considered in the food industry—and almost exclusively for risk assessment in areas such as heavy metal intake from seafood.[7,8] Most of the work presented in EVT literature tends to be more theoretical and mathematical, and the number of practical application stories is small. An exception is the excellent introduction by Geladi and Teugels[9] that provides an introduction to extreme value theory and suggested the methods could be useful in process characterization.

In EVT, an essential step is separating the extreme from nonextreme observations in a data set. The present manuscript will use the so-called generalized extreme value (GEV) distribution for the block maxima (BM; also known as the Gumbel method) approach in which data are divided into conceptually meaningful blocks, and from each block, the maximum Δ% (or, eg, some parameter minimum in other investigative scenarios) is collected to form a new array of data points to model on. As an alternative, the point over threshold (POT) methodology is applied where a threshold is determined and all values exceeding it, collectively over the full data record, are extracted and used for modeling. It is emphasized again that the tail or extreme values of many processes do not match well with the bulk distributions commonly used, and this was already recognized early on in the development of modern/present day statistical inference.[10] The history of EVT will not be covered here; for this, we refer to the recent review by Gomes & Guillou[11] plus references therein and the somewhat dated, but excellent, introduction by Geladi & Teugels.[9] The focus in this work will be on the practical aspects. Through real process data, we will demonstrate how the methods can be applied and interpreted and show the usefulness of characterizing extreme events in industrial food production, despite the sometimes challenging theoretical background/frameworks found in literature.

2 | BLOCK MAXIMA

Traditionally, BM has been used in, eg, hydrology where obvious periodicity or blocking is present because parameters such as annual extreme water levels are used. From these values, statistical likelihoods such as how high could the 100 year flood be (a return-level), or how frequent can a flood of this magnitude be expected (a return-time). As another example, in the field of wind engineering, yearly maxima have traditionally been used to estimate extreme wind speeds. To extract independent identically distributed (IID) extreme values, a blocking scheme called method of independent storms was introduced. Here, the fastest recorded wind is extracted from each storm defined by in-between lull periods (originally suggested by Cook in 1982).[12,13] In characterizing a sequence or time series originating from an industrial production as shown in Figure 1D, finding a blocking scheme is not so straightforward.

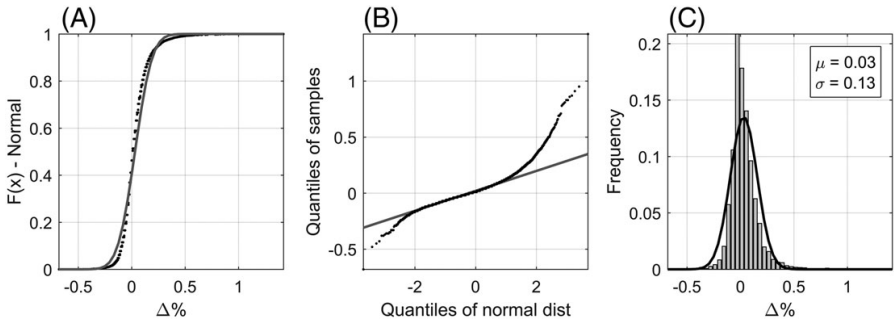

FIGURE 2 Normal distribution fitted to Factory 1a deviation in moisture content from set-point data (n = 4565); A, cumulative distribution function; B, QQ-plot; and C, probability density function

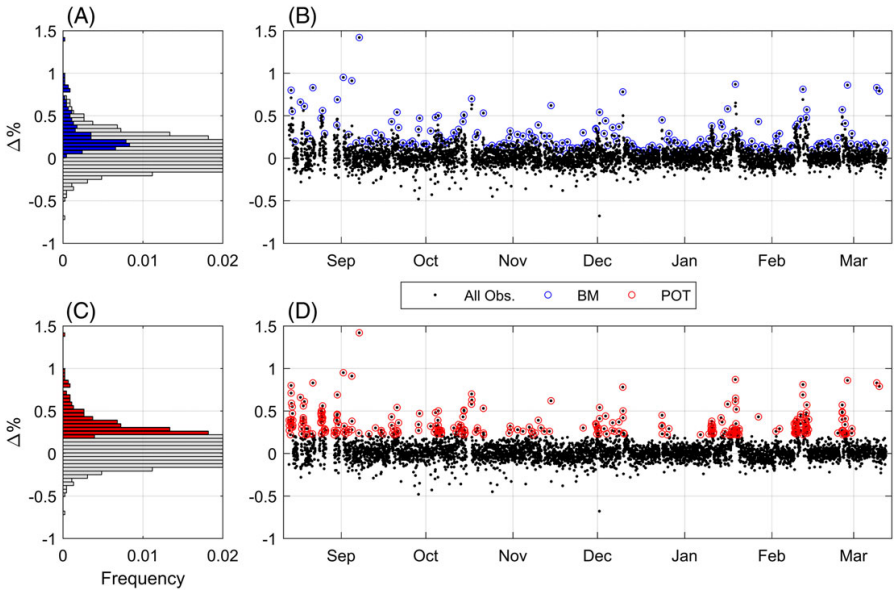

FIGURE 3 Histogram of extreme observations extracted from Factory 1a by the A, block maxima (BM) and C, point over threshold (POT) approach superimposed onto the histogram of all observations. Values extracted with B, BM ($k = 203$) and D, POT (k = 299) superimposed onto all data measured over time (n = 4565)

Food-processing equipment will be subjected to production stops due to regular cleaning, maintenance, and limited supply from upstream, rendering block size less obvious in this semicontinuous mode of operation. The disruptive food production process implies the need for a filtering approach to blocking since no fixed or natural periodicity exists as opposed to, eg, hydrology. Spray driers habitually run several days in a row, and production only stops for CIP or lack of feedstock. One such time segment we will call a production run for convenience although semicontinuous production is a more accurate description since there should ideally be no dynamic trends in between stops.[5] It is safe to assume that individual production runs are independent, and in our example, no explicit information about production stops was available. To systematically segment the process data into runs, we suggest and applied a filtering procedure. First, all appropriate production runs were identified based on two criterion: (1) more than 3 hours between two consecutive measurements marks a new run and (2) runs shorter than six hours are eliminated. During a continuous production period, the data were segmented every 24 hours to obtain more blocks inside the same production run. Implicitly, it was assumed that 24 hour blocks were sufficient in extracting independent observations of extreme events. The overall result of the extraction filtering can be seen in Figure 3A,B, where the average block size is 21.3 hours. The average length will later on be used to translate statistical findings back into production hours.

Formally, consider $\widetilde{\boldsymbol{X}}$ (n × 1) containing n observations of a parameter of interest, in our case, the full record of hourly deviation from the moisture set-point (Figure 1D). Split $\widetilde{\boldsymbol{X}}$ into k blocks—eg, 24 hour periods—and extract the maximum from each block into $\boldsymbol{X}$ (k × 1), thus containing k local maximum of Δ% that will be modeled on. It was shown by Fisher et al[10] in 1928 that the parent distribution F of the extreme values in $\boldsymbol{X}$ will have attractors (asymptotes) that belong to 1 of 3 distributions: Gumbel (Equation 3a; also known as the extreme value type I distribution), Fréchet (Equation 3b; type II) or reverse Weibull (Equation 3c; type III) with the respective CDFs[11] (see Figure 4):

$$\mathrm{F}(x) = \exp\left(-\exp\left(-\frac{x-b}{a}\right)\right), \xi = 0, \quad x \in \mathbb{R}, \tag{3a}$$

$$\mathrm{F}(x) = \exp\left(\left(-\frac{x-b}{a}\right)^{-1/\xi}\right), \quad \xi > 0, \quad x > 0 \quad | \quad x \leq 0, \quad F(x) = 0, \tag{3b}$$

$$\mathrm{F}(x) = \exp\left(-\left(-\frac{x-b}{a}\right)^{-1/\xi}\right), \xi < 0, \quad x > b \quad | \quad x \leq b, \quad F(x) = 1, \tag{3c}^{\dagger}$$

[†]The regular Weibull distribution is defined for minima only; the equation shown above is known in literature as the reverse Weibull distribution.

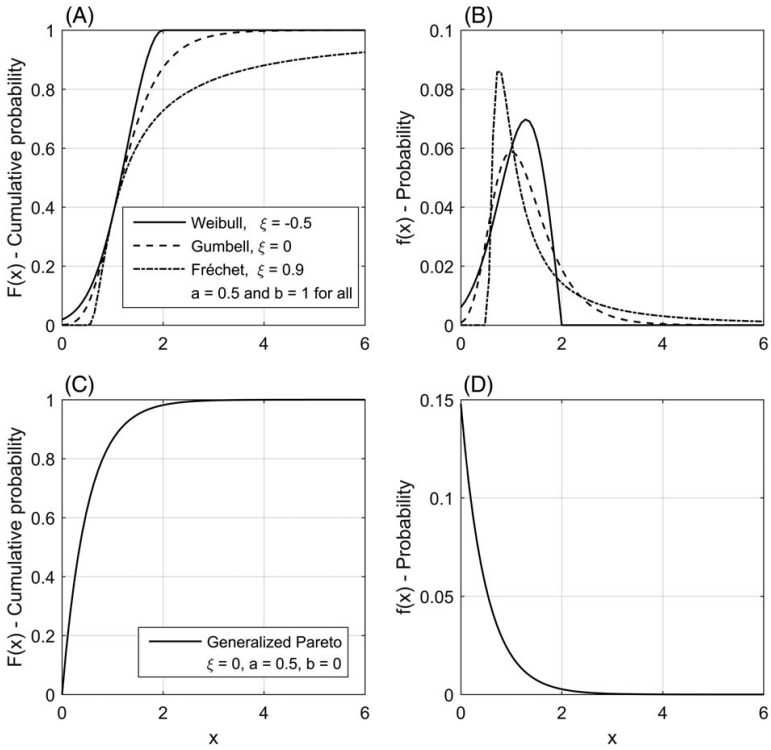

FIGURE 4 A, cumulative distributions function (CDF) and B, probability density functions (PDF) for block maxima distributions; C, CDF and D, PDF for point over threshold distribution

where b is the location parameter of the extreme value distribution, a is the scale or dispersion parameter, and ξ is the shape parameter or extreme value index. It is interesting to note the similarity between the three extreme value distributions (EVDs) in 3a and the limiting cases they represent. In a type III situation, the extremes are (upper) bounded; while for type II, the EVD tail is thick, and a wider range of extremes is predicted. The corresponding and slightly more involved PDFs (4a) can also be seen in Figure 4:

$$f(x) = \frac{1}{a}\exp\left(-\left(\frac{x-b}{a} + \exp\left(-\frac{x-b}{a}\right)\right)\right), \xi = 0, \quad x \in \mathbb{R}, \tag{4a}$$

$$f(x) = \frac{\xi}{a}\left(\frac{x-b}{a}\right)^{-1-\xi}\exp\left(-\left(\frac{x-b}{a}\right)^{-\xi}\right), \xi > 0, \quad x \geq 0, \tag{4b}$$

$$f(x) = \frac{\xi}{a}\left(\frac{x-b}{a}\right)^{\xi-1}\exp\left(-\left(\frac{x-b}{a}\right)^{\xi}\right), \xi < 0, \quad x > b. \tag{4c}$$

The 3 different distributions from 3a and 4a can be generalized (or combined) into the GEV distribution[11]:

$$GEV(\xi, x) = \begin{cases} \mathrm{F}(x) = \exp\left(-\left(1+\xi\frac{x-b}{a}\right)^{-1/\xi}\right), & \xi \neq 0, \quad 1+\xi\frac{x-b}{a} \geq 0, \\ \mathrm{F}(x) = \exp\left(-\exp\left(-\frac{x-b}{a}\right)\right), & \xi = 0, \quad x \in \mathbb{R}. \end{cases} \tag{5}$$

The shape parameter, ξ, determines the type of EVD (more Gumbel, Fréchet, or reverse Weibull like). If the shape parameter is estimated to be below 0, the data follows a Weibull-type distribution ($\xi < 0$), which is characterized by a finite tail, where the Gumbel ($\xi = 0$) and Fréchet ($\xi > 0$) distributions have infinite tails, with the latter being more heavy tailed (Figure 4B). This shape parameter is thus a key indicator to characterize the behavior of extremes in a production proces and gives the freedom to maneuver between the classical type I, II, and III EVDs. It is advisable to interpret the model parameter estimates at the end of an analysis because it might be beneficial to characterize the process performance of extreme values in terms using the three ξ architypes (3a and 4a).

The PDF for the GEV distribution is

$$GEV(\xi,x)=\begin{cases} f(x)=\frac{1}{a}\left(1+\xi\frac{x-b}{a}\right)^{-1-1/\xi}\exp\left(-\left(1+\xi\frac{x-b}{a}\right)^{-1/\xi}\right), & \xi\neq 0,\\ f(x)=\frac{1}{a}\exp\left(-\frac{x-b}{a}\right)\exp\left(-\exp\left(-\frac{x-b}{a}\right)\right), & \xi=0. \end{cases} \tag{6}$$

Using this distribution and the parameter estimates, the next obvious step is statistical inference on the data. The *return-time*, T, is the average frequency by which a value of level x can be expected, and it is obtained from the CDFs[12] in Equations 3 or 5:

$$F(x)=1-T^{-1}, \tag{7a}$$

$$T=(1-F(x))^{-1}. \tag{7b}$$

The *return-level* is the value that on average can be expected over a period T and can be estimated from the inverse of the CDF:

$$x=F^{-1}\left(1-T^{-1}\right)=b+\frac{a}{\xi}\left(1-\left(-\ln\left(1-T^{-1}\right)^{k}\right)\right),\quad \xi\neq 0, \tag{8a}$$

$$x=F^{-1}\left(1-T^{-1}\right)=b-a\,\ln\left(-\ln\left(1-T^{-1}\right)\right),\quad \xi=0. \tag{8b}$$

Parameters estimated in Equations 3, 4, 5 and 6 are in the unit "block." Each extreme observation represents the block from which it was extracted and thus the time span covered by the block. Consequently, the return-time estimated by the EVDs will be expressed of number of blocks, not production hours directly. This will be elaborated upon in the Section 4.

The BM extreme observations from Figure 3B are fitted to the generalized EVD from Equations 5 and 6, shown in Figure 5A-C, using a maximum likelihood estimation protocol (MATLAB® R2014b including the Statistics Toolbox™, Mathworks, USA).

3 | POINT OVER THRESHOLD

The POT approach is more decision free compared to BM. Rather than extracting the most extreme observation per block, and thus having to define the unit block, all values exceeding a defined threshold are extracted from data series $\widetilde{X}$. It is argued that with a properly defined threshold, more relevant extreme data are extracted this way, leading to more observations and a lower uncertainty in the parameter estimation. The extracted values can be modeled by the asymptotic generalized Pareto distribution (GPD):

$$GPD(\xi,x)=\begin{cases} F(x)=1-\left(1+\xi\frac{x-b}{a}\right)^{-1/\xi}, & \xi\neq 0, 1+\xi\frac{x-b}{a}\ x>0,\\ F(x)=1-\exp\left(-\frac{x-b}{a}\right), & \xi=0,\ \ x>0. \end{cases} \tag{9}$$

with the corresponding PDF

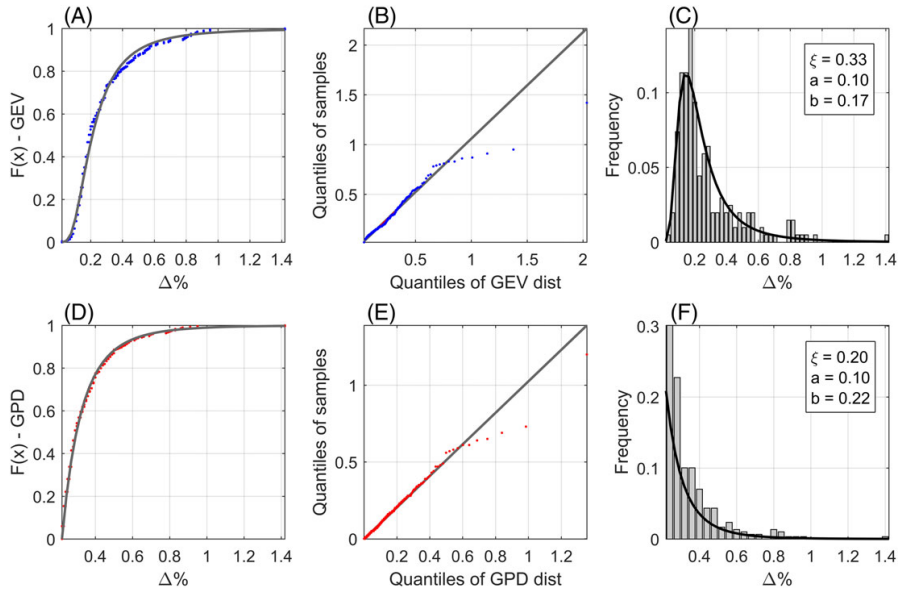

FIGURE 5 Generalized extreme value/GEV distribution (k = 203); A, cumulative distribution function (CDF); B, QQ-plot; and C, probability density function (PDF); generalized Pareto distribution (GPD) (k = 299) D, CDF, E, QQ-plot, and F, PDF

$$GPD(\xi,x)=\begin{cases} f(x)=\frac{1}{a}\left(1+\xi\frac{x-b}{a}\right)^{-1-\frac{1}{\xi}}, & \xi\neq 0, \\ f(x)=\frac{1}{a}\exp\left(-\frac{x-b}{a}\right), & \xi=0, \end{cases} \quad (10)$$

where ξ again is the shape parameter, a is the scale parameter, and b is the location parameter. The location parameter, b, however, is directly determined by the selected threshold. An example of the GPD can be seen in Figure 4. To estimate, eg, return-time T, the extraction procedure is assumed to behave according to a Poisson process with rate parameter $\lambda = k/n$, where k is the number of data points extracted (which is thus a function of the selected threshold) and n is the total number of observations—in our case, hourly moisture measurements.[12] From Equation 9 and the rate parameter λ, the return-time T for a value of level x using POT can be estimated[2]:

$$F(x)=1-(\lambda T)^{-1}, \quad (11a)$$

$$T=(\lambda(1-F(x)))^{-1}. \quad (11b)$$

The POT return-level can be estimated through the inverse CDF:

$$x=F^{-1}\left(1-(T\lambda)^{-1}\right)=b+\frac{a}{\xi}\left(1-(\lambda T)^{-\xi}\right), \quad \xi\neq 0, \quad (12a)$$

$$x=F^{-1}\left(1-(T\lambda)^{-1}\right)=b+a\ \ln(\lambda T), \quad \xi=0. \quad (12b)$$

While the advantage of POT is usage of typically more data points leading to better estimation of model parameters, the challenge is where to set the threshold value. Setting it too close to nonextreme values will bias the estimated parameters towards nonextreme or normal process behavior, while setting the threshold too excessive will lead to high uncertainty in the estimated parameters due to an insufficient number of data points—an example of the so-called bias-variance trade-off in model building. Despite threshold selection being the most crucial decision when using the POT method in EVT, no clear-cut strategy has been established.[14] A heuristic approach was chosen here by evaluating the fit in QQ-plots (similar to what is shown in Figure 2B) for different POT threshold choices. It was observed that the uncertainty in the estimate increases when lowering the threshold, as a result of the bias-variance trade-off described previously. The shape parameter ξ did not fluctuate much and remained close to 0.1. The threshold value for further evaluation was set to be 0.22% moisture difference between process set-point and measured hourly value. The parameter estimation and distribution fitting using this threshold and GPD is presented in Figure 5D-F.

It should be recognized here that the POT strategy is more influenced by, eg, steady-state behavior/detrending of the data record. An exaggerated example could be if we apply the POT approach to the (raw) data shown in Figure 1A, which illustrates a set-point change. If a fixed threshold is applied, only extreme data points from the second phase would be extracted, masking the extreme production values from the first part. This way, the assumption of IID observations ending up in the EVT data record is violated. The BM method on the other hand always (forcibly) samples from the entire data length.

4 | INTERPRETATION OF THE MODELS

Figure 5 shows the CDF, QQ-plot, and PDF for the two EVT strategies (BM and POT) applied to the moisture deviation data record from Factory 1a. For comparison, the classical normal distribution theory, applicable to the bulk data, is seen in Figure 2. From the QQ-plot of the normal distribution, it is clear that the positive (and negative) tail of the data collection is not modeled effectively, while for both the BM and POT method, the extremes are modeled well. Selected return-times T for Factory 1a are presented in Table 1. Note that the return-time for BM (Equation 7b) comes of blocks, not directly production hours. To make the results between BM and POT comparable and make the finding more suitable for daily production practice, the BM return periods are multiplied with the average block time span (21.3 hours). The smallest exceedance level of 0.2% from the set-point cannot be determined for the POT model since the threshold was set at 0.22%. The normal distribution return-time estimates are sensible for an exceedance of 0.2%, but at 0.4%, they significantly deviate from the EVT-based estimates, and for an exceedance of more than 0.5%, no meaningful result can be obtained under the normal distribution assumption.

The results in Table 1 also highlight a distinction between BM and POT. Taking the return-times for 0.4%, for example, the POT analysis in combination with all hourly observations from 1 production campaign of 8 months (Figure 1D) predicts a deviation of 0.4% on average every 67 production hours. It does not take into consideration any systematic variation from factors such as seasonal weather influences, time-after-last-maintenance, etc, which might lead to extreme values occurring more often than normal over certain intervals, hence lumped together. Block maxima and its 24-hour filtering approach on the other hand implicitly constrain the influence of some of these external aspects on the extreme values. The same 0.4% deviation case for BM should thus

TABLE 1 Return-times in production hours for selected exceedance levels for the normal distribution, EVT-BM, and EVT-POT; 95% CIs estimated from Naïve bootstrapping[15]

Δ%	Normal Distribution Return-Time, h	BM Return-Time, h	POT Return-Time, h
0.2%	12 [11-14]	39 [36-45]	-
0.3%	71 [54-95]	73 [62-92]	31 [28-35]
0.4%	753 [477-1221]	129 [103-174]	67 [56-80]
0.5%	$1.51 \cdot 10^4$ [$0.72 \cdot 10^4$-$3.21 \cdot 10^4$]	212 [161-305]	129 [105-171]
0.6%	$53 \cdot 10^4$ [$19 \cdot 10^4$-$160 \cdot 10^4$]	327 [227-503]	234 [179-343]

Abbreviations: BM, block maxima; CIs, confidence intervals; EVT, extreme value theory; POT, point over threshold; -, not available/not defined

be read: with all hourly observations from 1 production campaign of 8 months, using the 24-hour filtering strategy, we predict a deviation of 0.4% on average ever 129 production hours. The latter might initially seem overly complex, but the two are complementary.

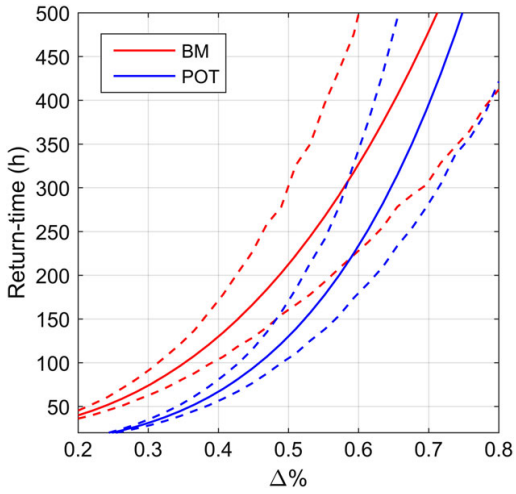

FIGURE 6 Return-times versus exceedance from set-point for moisture contents for Factory 1a; estimated Return-times (solid line) and 95% CIs (dashed) estimated from naïve bootstrapping.[15] BM, block maxima; CIs, confidence intervals; POT, point over threshold

In Table 1, the discrepancy between BM and POT results for EVT analysis decreases as the exceedance from the set-point increases (while at the same time leading to wider uncertainty intervals, as expected). To visualize this pattern, Figure 6 shows the exceedance from set-point as a function of return-time for BM and POT.

Table 2 shows the EVT parameter estimation results for all 3 factories. From the BM method, the shape parameter can be observed to be above 0 for Factory 1a,b, while Factory 2 has a shape parameter confidence interval that includes 0. This suggests that the extreme observations in Factory 1a and 1b follow a Fréchet/type II distribution, which is the heavy-tailed infinite profile (Figure 4B), while the extreme observations in Factory 2 follow a Gumbel/type I distribution with the infinite, but not so heavy tail profile (Figure 4B). This information supports the estimated return-levels above that indicate Factory 1a and Factory 1b are more prone to produce milk powder with a larger deviation from the moisture set-point given the same production time.

Many approaches can be taken to use the information captured in the EVT models. As an example, Table 3 describes the expected return-level after 720 hours of production for the three different production lines. As expected, Factory 1a and Factory 1b have almost identical estimates since they are fed from the same spray drier. Factory 2 on the other hand has a significantly lower return-level, telling us that over a period of 720 hours, a much smaller worst case extreme deviation can be expected.

TABLE 2 Modeling choices and estimated extreme value distribution parameters for the 3 production lines; 95% CI estimated from bootstrapping

	Factory 1a	Factory 1b	Factory 2
BM	Avg. block size: 21.3 ξ = 0.330 [0.238-0.500] a = 0.098 [0.081-0.113] b = 0.173 [0.159-0.188]	Avg. block size: 21.3 ξ = 0.388 [0.287-0.496] a = 0.092 [0.079-0.105] b = 0.151 [0.138-0.165]	Avg. block size: 14.7 ξ = 0.037 [−0.077-0.148] a = 0.121 [0.108-0.133] b = 0.179 [0.160-0.201]
POT	b = 0.22 Δ % λ = 0.0655 ξ = 0.199 [0.050-0.342] a = 0.105 [0.085-0.127]	b = 0.22 Δ % λ = 0.0519 ξ = 0.098 [−0.072-0.240] a = 0.144 [0.115-0.182]	b = 0.22 Δ % λ = 0.0509 ξ = -0.099 [−0.255-0.018] a = 0.144 [0.121-0.175]

Abbreviations: BM, block maxima; CI, confidence interval; POT, point over threshold.

TABLE 3 Estimated return-levels for moisture exceedance during 720 hours (30 days) of production for the 3 production lines; 95% CI estimated from Naïve bootstrapping[15]

	Factory 1a (Δ%)	Factory 1b (Δ%)	Factory 2 (Δ%)
BM	0.82 [0.68-1.06]	0.84 [0.67-1.05]	0.68 [0.59-0.78]
POT	0.83 [0.71-0.97]	0.85 [0.74-0.97]	0.66 [0.59-0.72]

Abbreviations: BM, block maxima; CI, confidence interval; POT, point over threshold.

5 | MICROBIOLOGICAL DATA CASE

To illustrate the diversity of EVA/EVT applications in the food industry, a second example deals with biological counts in whole milk powder. These counts are determined in a quality assurance laboratory a variable number of times for one batch of produced powder. Data from July 2013 to February 2016 are available from two factories, here called J and K. The batch size and processing speed ($kg \cdot hour^{-1}$ or an equivalent) differs from site to site, and the *timescale* is expressed in number of produced batches over a time period. Biological counts have been masked due to confidentiality reasons. For this case, only the BM approach is used since we expect count observations within 1 batch to be correlated, and to get rid of this signal correlation, a blocking strategy can be used. Two blocking extreme value strategies were applied: (1) pool all observations during a 10 day period and (2) pool all observations from 10 consecutive batches. Figure 7 shows counts obtained from Factory J over 2 full and 1 partial production season. For Factory J, the 10-day blocking strategy and the 10-batch blocking strategy are quite different. Factory K (data not shown) follows a more constant production schedule, leading to more similar blocks for the two strategies.

Generalized extreme value models were fitted to the data extracted with the 10-day and 10-batch blocking strategies for both factories J and K. The PDFs and QQ-plots can be seen in Figure 8. From the QQ-plots in Figure 8A-B, it follows that the fit for Factory J differs somewhat between the strategies—this is also observed in Figure 8C where the PDFs do not overlap completely. However, when estimating return-times for the biological counts higher than level B, the results converge. For Factory K, the two strategies only result in very slight differences in the model fits (Figure 8D-E) leading to almost completely overlapping PDFs (Figure 8F). Comparing the 10-batch strategy between the two factories (Figure 8G), it is obvious that they do not have the same extremal behavior; the same is observed for the 10-day strategy in Figure 8 H. The conclusion is that the blocking strategy is less important as long as it is reasonable and that EVT can be used to compare the two factories biological counts extremal behavior, illustrating the versatility of applying EVT on production and process data.

6 | CONSIDERATIONS IN EXTREME VALUE ANALYSIS

The natural focus in EVT is in, eg, hydrology extreme flooding or other low frequency events estimated via extrapolation from the available sparse data. In the characterization of food production systems on the other hand, extremes with a moderate frequency might be of economic significance, while extrapolated extremes such as an unrealistically high-moisture value in milk powder are prevented by the automated controls and operator interventions. From this perspective, production characterization as discussed in this investigation can be seen as operating on the left-hand side of

FIGURE 7 Biological counts for 2 full and 1 partial season; blocks illustrate 10-day and a 10-batch pooling strategy; y-axis is in arbitrary and truncated biological counts for confidentiality reasons

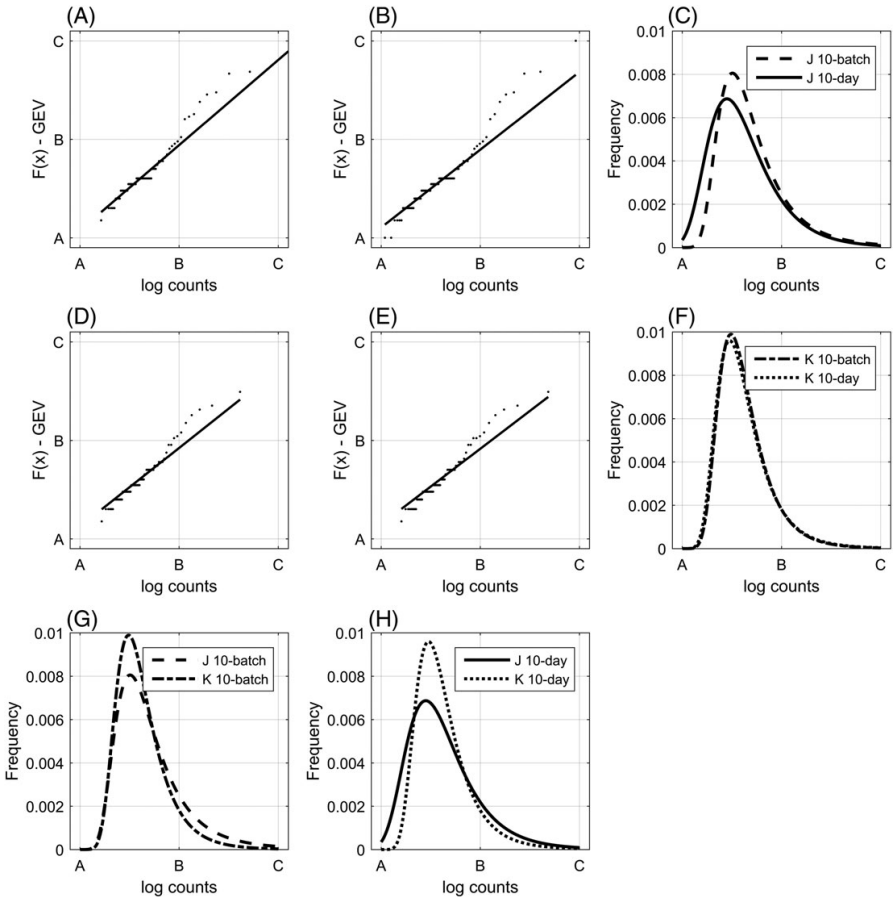

FIGURE 8 Overview of model fit across strategies and factories: A, Factory J 10-batch strategy; B, Factory J 10-day strategy; C, probability density functions (PDFs) for Factory J; D, Factory K 10-batch strategy; E, Factory K 10-batch strategy; F, PDFs for Factory K; G, PDFs for 10-batch strategy across factories; and H, PDFs for 10-day strategy across factories

Figure 6 while flooding is naturally positioned on the right-hand side (out of necessity).

A potential limitation with the BM strategy in extreme value analysis is not using all extreme observations available as compared to the POT approach. However, Ferreira and De Haan[16] argue that there are several practical reasons for using the BM method: often only BM values (such as monthly or yearly maxima) are available, and the method may thus be easier to apply as block periods appear naturally. Moreover, if the observations in the data record are not IID, the POT method may not be appropriate. The last argument is relevant in process monitoring. The rapid collection of data from a production process often results in high autocorrelations in the signal; implicitly meaning that POT extracted values may not be independent. Setting the threshold for the POT method may alter the dependency (correlation) between the extracted data but not necessarily solve the problem.

This threshold is a parameter that must be defined by the user, and yet, no clear-cut method has been established.[14] Similarly, how to select the best blocking scheme when using the BM method is not readily obvious. In this case, however, one can use domain knowledge as guideline to set up an appropriate plan. We note that return-time estimates at a return-level 0.5 % for our moisture values in milk powder example did not change significantly as a function of either threshold selected (tested from b = 0.1% to 0.4%[‡]) or blocking size (from 10 to 40 hours[§]) for, respectively, the POT or BM approach.

[‡]Threshold, return-time for 0.5 % deviation from set-point for Factory 1a: 0.10% (RT (h) = 123 [99 - 158]), 0.22% (129 [105 - 171]), 0.40% (119 [99 - 146]).

[§]Block size, return-time for 0.5 % deviation from set-point for Factory 1a: 10 hours (RT (h) at 0.5 % = 204 [152 – 300]), 24 hours (212 [161 – 305]), 40 hours (223 [171 - 307]).

An advantage of the BM approach is using the shape parameter to characterize the processing system. In the moisture case, factories 1a and 1b's extreme observations suggested the Fréchet or type II extremal distribution inferring a heavily tailed pattern, while Factory 2's extreme observations showed a Gumbel distribution with a less heavy tail pattern. In the case of industrial process characterization where several observations per batch are available, we are inclined to recommend using the BM method with a suitable blocking scheme. Despite the fact that not all extreme observation may be included, the ones that are included are more likely to obey the assumptions of the models.

Care should be taken concerning the sampled signal when applying extreme value theory. Extreme value theory requires long series of observations to contain a reliable amount of extreme data points from a stable, in control system. In a production environment, this can prove difficult as, eg, continuous process improvements (hence, permanent optimization and tuning) may perturb the system. For example, if the analytical method changes, ie, the NIRS calibrations are updated or adjusted and this changes the uncertainty in predictions, this new data cannot necessarily be combined with old data. Or, if the manufacturing practice changes, the system may also change, thus invalidating the models built on historic records. This is true for all statistical models, but since EVT works on extremes, often collected over long time periods, these models are more at risk than most others.

In the cases presented here, extreme events occur over time, as explicitly expressed in the equations. However, time T can be converted to a simple counting probability. An example could be measuring the active pharmaceutical ingredient in many pills from 1 true batch process with NIR or Raman spectroscopy. From such data, the return-time or, more relevant, the probability of producing pills with a too low or high API contents could be estimated in the unit one-in-so-many pills for example.

It is important to note that alternative approaches incorporating the full data set when inferring on the extreme values are possible. These modeling strategies will likely be more challenging (in model selection, statistical skills of the modeler, more parameters to estimate, etc) and should also take into account autocorrelation where relevant, but should be considered as alternatives to EVT.

Obtaining knowledge regarding the expected frequency of a specific exceedance, or size of an exceedance during a certain time period, can be used to uncover risk. Combining this knowledge with production costs can help focus and identify future improvements or savings. This decision is most likely unique to each the specific manufacturer, but regardless of whether an unknown risk or need for improvement is identified, it will at least be known. Despite EVT research being dominated by—in our opinion—overly complicated theory and abstract concepts, the tools can successfully be applied in process characterization. We hope that this tutorial will make EVT more accessible in both research and daily industrial practice.

ACKNOWLEDGEMENTS

We would like to thank the people at Fonterras Research and Development Centre, especially Rob Crawford and Roger Sterling.

This work was supported by the REuse of WAteR in the food and bioprocessing inDustries (REWARD) consortium funded by the Danish Council for Strategic Research, Programme Commission on Health Food and Welfare.

REFERENCES

1. van den Berg F, Lyndgaard CB, Sørensen KM, Engelsen SB. Process analytical technology in the food industry. *Trends Food Sci Technol*. 2013;31(1):27–35. https://doi.org/10.1016/j.tifs.2012.04.007
2. Abild J, Andersen EY, Rosbjerg D. The climate of extreme winds at the great belt, Denmark. *J Wind Eng Ind Aerodyn*. 1992;41(1–3):521–532. https://doi.org/10.1016/0167-6105(92)90458-M
3. Minkah R. An application of extreme value theory to the management of a hydroelectric dam. *Springerplus*. 2016;5:96 https://doi.org/10.1186/s40064-016-1719-2
4. Khalifa M, Khan F, Haddara M. A methodology for calculating sample size to assess localized corrosion of process components. *J Loss Prev Process Ind*. 2012;25:70–80. https://doi.org/10.1016/j.jlp.2011.07.004
5. Holroyd SE, Prescott B, Mclean A. The use of in- and on-line near infrared spectroscopy for milk powder measurement. *J Near Infrared Spectrosc*. 2013;21:441–443. https://doi.org/10.1255/jnirs.1073
6. Chavez-Demoulin V, Davison AC. Modelling time series extremes. *Revstat Stat J*. 2012;10(1):109–133.
7. Maulvault AL, Cardoso C, Nunes ML, Marques A. Risk-benefit assessment of cooked seafood: black scabbard fish (*Aphanopus carbo*) and edible crab (*Cancer pagurus*) as case studies. *Food Control*. 2013;32(2):518–524. https://doi.org/10.1016/j.foodcont.2013.01.026
8. Tressou J, Crépet A, Bertail P, Feinberg MH, Leblanc JC. Probabilistic exposure assessment to food chemicals based on extreme value theory. Application to heavy metals from fish and sea products. *Food Chem Toxicol*. 2004;42(8):1349–1358. https://doi.org/10.1016/j.fct.2004.03.016
9. Geladi P, Teugels JL. Extreme value theory and its potential role in chemometrics: an introduction. *J Chemometr*. 1996;10:547–567. https://doi.org/10.1002/(SICI)1099-128X(199609)10:5/6<547::AID-CEM473>3.0.CO;2-B
10. Fisher RA, Tippett LHC. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math Proc Cambridge Philos Soc*. 1928;24(2):180–190. https://doi.org/10.1017/S0305004100015681
11. Gomes MI, Guillou A. Extreme value theory and statistics of univariate extremes: a review. *Int Stat Rev*. 2015;83(2):263–292. https://doi.org/10.1111/insr.12058

12. Palutikof JP, Brabson BB, Lister DH, Adcock ST. A review of methods to calculate extreme wind speeds. *Meteorol Appl*. 1999;6:119–132. https://doi.org/10.1017/S1350482799001103
13. Harris RI. Improvements to the 'method of independent storms'. *J Wind Eng Ind Aerodyn*. 1999;80(1):1–30. https://doi.org/10.1016/S0167-6105(98)00123-8
14. Beirlant J, Caeiro F, Gomes MI. An overview and open research topics in statistics of univariate extremes. *Revstat Stat J*. 2012;10(1):1–31.
15. Wehrens R, Putter H, Buydens LMC. The bootstrap: a tutorial. *Chemom Intel Lab Syst*. 2000;54(1):35–52. https://doi.org/10.1016/S0169-7439(00)00102-7
16. Ferreira A, De Haan L. On the block maxima method in extreme value theory: PWM estimators. *Ann Stat*. 2015;43(1):276–298. https://doi.org/10.1214/14-AOS1280

How to cite this article: Skou PB, Holroyd SE, van den Berg F. Tutorial – applying extreme value theory to characterize food-processing systems. *Journal of Chemometrics*. 2017;e2896 https://doi.org/10.1002/cem.2896

5.2 Estimating uncertainty

In **Paper III** confidence intervals were estimated via the naïve bootstrapping approach as suggested by Naess and Clausen.[94] The bootstrap relies on the assumption that the uncertainty of a statistical parameter (here the return-time or return-level) can be estimated from the variability generated by calculating the statistic numerous times from new datasets constructed through random re-sampling (with replacement) of the original dataset. This is known as the *plug-in* principle.[95] The bootstrap samples for the food processing case were constructed by re-sampling the extracted extreme observations into new datasets and fitting the GEV distribution, estimating either the return-time or return-level and collecting these bootstrap estimates. In total 1000 bootstrap samples (and estimates) were calculated. From these estimates the 2.5% and 97.5% confidence limits could be found for the three parameters a, b and ξ.

However, during calculation of the confidence intervals for the manuscript, meaningless result were often obtained from a given bootstrap sample. This was a consequence of the models not converging to meaningful solutions, leading to estimates of size *infinite* or simply not converging leading to *not-a-number* results. Investigating this issue it became apparent that some bootstrap samples were chosen *awkwardly*, generating datasets not resembling those of the original dataset. Due to the asymptotic nature of the extreme value distributions, this can lead to estimates that are e.g. outside the numerical limits of Matlab ($> 10^{16}$), or situations were it is impossible to fit the data sufficiently well.

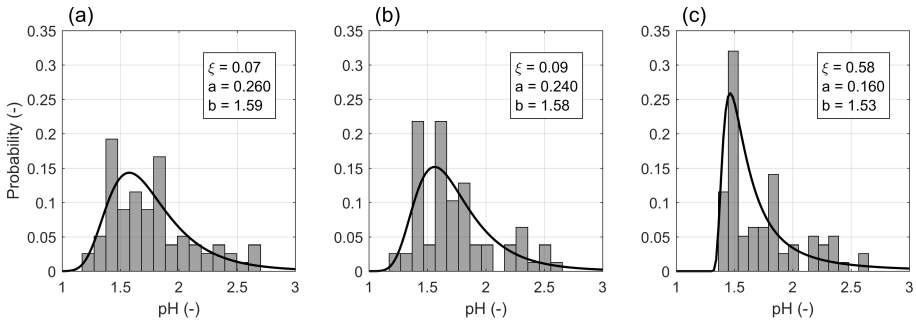

Figure 5.8: Examples of good (b) and problematic (c) bootstrap sample compared to the original data set (a) derived from the laboratory experiment.

Although failure to converge happened very seldom in the analysis of the pH process signal data (Figure 5.2), poor bootstrap samples could be found. Figure 5.8 shows the original data set and an example of an unluckily / unfortunate (defined by an estimated return-level larger than 8) and well-chosen bootstrap sample. The tail-parameter (ξ) can be seen to differ considerably from both the

good bootstrap sample and the original data set model's tail-parameter. Of course the purpose of bootstrapping is to find the boundaries, but due to the asymptotic nature of the distributions this is not always feasible.

For the publication this issue was circumvented by simply removing the non-numerical results obtained from the bootstrap procedure and finding the limits from this reduced set of bootstrap estimates. Circumventing this issue, however, is not a solution to the problem. How the problem ideally should be solved is left for further research, but the most promising strategy could be to switch from naïve/non-parametric to a semi-parametric tactic where the residuals rather than the observations are re-sampled.

5.3 Future applications of EVT in food processing

In EVT extremes are extracted from data records. In the case of Abild et al.[88] and the laboratory experiment presented in this chapter full data records were available with the aim of finding extreme observations. In the process data in **Paper III** records were available, however the goal of these measurements was not to identify extreme, but to monitor the process. The extremes could be modelled in the latter case as an added benefit of the monitoring strategy despite not being the original intention. The situation is different for e.g. the pipe corrosion scenario, the long-term study in Section 2.2 and the evaporator condensate characterisation investigations in Chapter 4 — here samples must be chosen / sampling must be done with the explicit aim of finding extreme observations. While pipe-corrosion problems are spatial and process characterisation are temporal, they still share the burden of choosing when / where to sample, or put differently, a sampling strategy is needed in these cases. In future novel applications of EVT it will probably be helpful to distinguish which scenario an application fits.

In the opening of this chapter the microbiological worst-case scenario for membrane permeate was outlined as an application that EVT could help solve. This would be a *pipe-corrosion* type of problem, where several production runs should be sampled a number of times similar to what was done in the long-term monitoring investigation. From these individual subsamples (production runs) the maxima could be extracted, and the maximum expected lactose concentrations inferred via the extreme value distributions. In order to get reliable estimates maybe up to 20 production runs should be sampled say five times each (speculative values). However this is not a straight forward method and is linked to active research in the pipe-corrosion estimation community.[90] This maximum estimated lactose concentration could then serve as input to the optimal growth medium / worst-case growth scenario in this type of process-water. This will aid in generating trustworthy microbiological assessment of process-water.

It should be noted that this worst-case scenario is only valid for the specific processing plant and not necessarily for any other plant running a similar process for several reasons including variations in feed composition and operating conditions such as temperatures and pressures, the physical organisation of the membranes, the age of the membranes, etc. In this regard any processing plant should ideally be characterised in its own right as many factors weigh-in on e.g. process-water quality. Due to deterioration and ageing of the system they should also be characterised periodically to ensure compliance.

The potential applications of EVT in food production are many. In addition to extreme deviations of key quality parameters and extreme microbial loads in the final product many more examples could be imagined such as the deviation of declared and actual weight of packaged products, characterizing and comparing milk suppliers with regards to fat, protein, lactose, somatic cell count or acidification speed for cheese. In principle any key quality attribute that is measured can potentially be described with EVT. In this sense it should be considered complementary to other more often calculated parameters such the mean, the spread around the mean, and autocorrelation for dynamic behaviour, thus providing additional information.

Chapter 6

Conclusions and perspectives

6.1 Conclusions

In the explorative work of this thesis, urea was found to be the main organic compound permeating RO membranes. Other larger organic molecules, including lactose, long-chain fatty acids, glycerol-phosphate and glutamic acid, were also found to permeate through in low concentrations. Element concentrations were overall very low as indicated by conductivity measurements, and co-varied closely as shown by ICP-OES quantification.

Urea permeates the RO membranes due to its similarity to water and thus constitutes a special case compared to other compounds such as lactose. Regardless, it was by far the most dominant organic molecule and therefore interesting to quantify in the process-water stream. NIRS was applied to quantify urea in laboratory scale and due to the relative low sensitivity of NIRS thorough uncertainty estimation was needed.

Despite the low nutrient level found in the membrane permeate, some micro organisms isolated upstream in the process were able to grow in the permeate. In order to establish appropriate holding times for this process-water stream worst-case scenarios should be set-up. In order to describe worst-case scenarios a novel statistical method, EVT, was explored and applied to a food processing system (not directly related to process-water). The results suggest that EVT could be of interest in microbial growth experiments.

Finally, evaporator condensate was tentatively characterised with a broad range of analytical techniques. Though results indicate aromatic amino acids, e.g. some protein source, was systematically present in the evaporator condensates. However, it was not possible to further identify the compound(s). It is suggested that additional analytical techniques such as SPME based GC-MS be used to capture volatile compounds to get a better understanding.

6.2 Perspectives

In order to consolidate the use and storage of process-water future studies on membrane permeate could include stratified sampling as discussed in section 5.3 in defining a worst-case scenario for microbial growth. This scenario should act as the growth medium for microbiological growth studies to evaluate allowable holding times.

To verify whether all carbon has been accounted for in the process-water stream, the approach by Vourch et al.[31] for dairy process-water and Busetti et al.[12] for wastewater where quantified compounds are compared to TOC or DOC could be implemented. Hence, process water could be analysed for TOC and comparing this to theoretical TOC contributions by the identified main organic compound (in this case urea in ROP membrane permeate). This will verify whether urea is indeed the main organic compound and help clarify whether any compounds were overlooked.

New analytical methods such as SBSE GC-MS should be considered as an alternative to the silylation derivatisation based GC-MS presented in this work. The evaporator condensate study indicated that volatile compounds were overlooked and this could also be the case for membrane permeates.

The present PhD project has dealt with the potential *use* of reclaimed process-water. In this scenario water is reclaimed from the raw material and used for a given purpose and then discharged. This relieves the immediate water-stress on the factory and is the foundation for further work. In Denmark the Danish Agriculture & Food Council, is working on a dairy industry-wide exemption (*branchekode* in Danish) from current legislation to allow the use of water fit-for-purpose rather than only distinguishing potable and non-potable water. In this framework effluents and process-waters of different qualities are distinguished and used appropriately. This will hopefully help producers to use lower quality water streams in lieu of potable water when appropriate. A system where process-water is segmented into different qualities with distinct storage life could be imagined. Process-water could be classified into e.g. high quality process-water which can sit for several days, intermediate quality which should be used within 24 hours and low quality which is discarded immediately.

This thesis work is part of the REWARD project which in the full title states that process-water is intended for *reuse*. In REWARD the vision is to make the the food and bioprocessing industry self-sustainable.[9] Self-sustainable factories — or closed-loop factories to use the terminology from circular economy — will be required to *reuse* process-water. While this work has focussed on the direct quality of process-water in a reuse scenario the long-term quality must also be taken into consideration. Build-up of persistent chemicals may become an issue. However,

currently process-water is not yet being reused in the dairy industry (or any other industry to the best of the authors knowledge).

The studies presented in the thesis have been based on processes at AFI processing facility in Nr. Vium (DK) where protein and lactose based ingredients are produced. If the principles studied here are to be rolled out in e.g. cheese production facilities, for using process-water reclaimed from whey, additional considerations should be taken. The starter cultures responsible for the acidification and flavour development in cheese production are prone to bacteriophage (virus) infections.[96] Bacteriophage infections are responsible for delayed and failed acidification and one suggested to inflict large variability into the final product with great economic consequences.[97,98] Bacteriophages propagates during acidification and are released into the water phase and are often found in high concentrations in the whey. Viruses are known to pass RO membranes (mainly due to some loss of integrity) and are also very challenging to monitor in real-time. Great care should be taken before implementing the results presented in this body of work. Simply put, the specific scenario and end-use should be considered carefully before implementing the use of process-water a food production process.

Bibliography

[1] United Nations (2017). Water. Accessed: 2017-07-06. URL: http://www.un.org/en/sections/issues-depth/water/index.html.

[2] World Health Organization (2017). Drinking-water. Accessed: 2017-07-06. URL: http://www.who.int/mediacentre/factsheets/fs391/en/.

[3] European Commision (2016). Eu resource efficiency scoreboard 2015. Technical report, European Commision. URL: http://ec.europa.eu/environment/resource_efficiency/targets_indicators/scoreboard/pdf/EU%20Resource%20Efficiency%20Scoreboard%202015.pdf.

[4] Rad, S. J. and M. J. Lewis (2014). Water utilisation, energy utilisation and waste water management in the dairy industry: A review. *International Journal of Dairy Technology 67*(1), 1–20. doi:10.1111/1471-0307.12096.

[5] Statistics Denmark. http://statistikbanken.dk/VANDRG2. accessed 2017-10-20.

[6] Meneses, Y. E., J. Stratton, and R. A. Flores (2017). Water reconditioning and reuse in the food processing industry: Current situation and challenges. *Trends in Food Science and Technology 61*, 72–79. doi:10.1016/j.tifs.2016.12.008.

[7] Walstra, P., J. T. M. Wouters, and T. J. Geurts (Eds.) (2006). *Dairy Science and Technology* (2nd ed.). Boca Raton, FL, USA: Taylor & Francis.

[8] Gekas, V. (1988). Terminology for pressure-driven membrane operations. *Desalination 68*(1), 77–92. doi:10.1016/0011-9164(88)80045-6.

[9] Holst, H. H., F. van den Berg, and S. B. Engelsen (2014). Water reuse and saving in the food industry. *New food 17*, 7–11.

[10] Pouliot, Y. (2008). Membrane processes in dairy technology-from a simple idea to worldwide panacea. *International Dairy Journal 18*(7), 735–740. doi:10.1016/j.idairyj.2008.03.005.

[11] Wenten, I. G. and Khoiruddin (2016). Reverse osmosis applications: Prospect and challenges. *Desalination 391*, 112–125. doi:10.1016/j.desal.2015.12.011.

[12] Busetti, F., M. Ruff, and K. L. Linge (2015). Target screening of chemicals of concern in recycled water. *Environ. Sci.: Water Res. Technol. 1*(5), 659–667. doi:10.1039/C4EW00104D.

[13] Nielsen, W. K. (Ed.) (2000). *Membrane filtration and related molecular separation technologies*. APV Systems.

[14] Bylund, G. (1995). *Tetra Pak Dairy Processing Handbook*. Tetra Pak Processing Systems AB.

[15] Ozaki, H. and H. Li (2002). Rejection of organic compounds by ultra-low pressure reverse osmosis membrane. *Water Research 36*(1), 123–130. doi:10.1016/S0043-1354(01)00197-X.

[16] Lee, S. and R. M. Lueptow (2001). Reverse osmosis filtration for space mission wastewater: membrane properties and operating conditions. *Journal of Membrane Science 182*(1-2), 77–90. doi:10.1016/S0376-7388(00)00553-6.

[17] Bridgewater, L., E. W. Rice, R. B. Baird, A. D. Eaton, and L. S. Clesceri (Eds.) (2012). *Standard Methods For The Examination of Water and Wastewater* (22nd ed.). Washington, DC, USA: American Public Health Association.

[18] Yip, V., S. Arntfield, and A. Hydamaka (1996). Potential for stainless steel microfiltration processing to reduce effluent from a fluid milk and ice cream processing plant. *Journal of Dairy Science 79*(4), 710–716. doi:10.3168/jds.S0022-0302(96)76417-2.

[19] Chmiel, H., V. Mavrov, and E. Bélières (2000). Reuse of vapour condensate from milk processing using nanofiltration. *Filtration and Separation 37*(3), 24–27. doi:10.1016/S0015-1882(00)88493-1.

[20] Mavrov, V. and E. Bélières (2000). Reduction of water consumption and wastewater quantities in the food industry by water recycling using membrane processes. *Desalination 131*(1-3), 75–86. doi:10.1016/S0011-9164(00)90008-0.

[21] Koyuncu, I., M. Turan, D. Topacik, and A. Ates (2000). Application of low pressure nanofiltration membranes for the recovery and reuse of dairy industry effluents. *Water Science and Technology 41*(1), 213–221.

[22] Balannec, B., G. Gésan-Guiziou, B. Chaufer, M. Rabiller-Baudry, and G. Daufin (2002). Treatment of dairy process waters by membrane operations for water reuse and milk constituents concentration. *Desalination 147*, 89–94. doi:10.1016/S0011-9164(02)00581-7.

[23] Nguyen, M., N. Reynolds, and S. Vigneswaran (2003). By-product recovery from cottage cheese production by nanofiltration. *Journal of Cleaner Production 11*(7), 803–807. doi:10.1016/S0959-6526(02)00130-0.

[24] Turan, M. (2004). Influence of filtration conditions on the performance of nanofiltration and reverse osmosis membranes in dairy wastewater treatment. *Desalination 170*(1), 83–90. doi:10.1016/j.desal.2004.02.094.

[25] Akoum, O., M. Y. Jaffrin, L. H. Ding, and M. Frappart (2004). Treatment of dairy process waters using a vibrating filtration system and NF and RO membranes. *Journal of Membrane Science 235*, 111–122. doi:10.1016/j.memsci.2004.01.026.

[26] Vourch, M., B. Balannec, B. Chaufer, and G. Dorange (2005). Nanofiltration and reverse osmosis of model process waters from the dairy industry to produce water for reuse. *Desalination 172*(3), 245–256. doi:10.1016/j.desal.2004.07.038.

[27] Suárez, E., A. Lobo, S. Álvarez, F. A. Riera, and R. Álvarez (2006). Demineralization of whey and milk ultrafiltration permeate by means of nanofiltration. *Desalination 198*(1-3), 274–281. doi:10.1016/j.desal.2007.11.087.

[28] Sarkar, B., P. P. Chakrabarti, A. Vijaykumar, and V. Kale (2006). Wastewater treatment in dairy industries - possibility of reuse. *Desalination 195*(1-3), 141–152. doi:10.1016/j.desal.2005.11.015.

[29] Frappart, M., O. Akoum, L. H. Ding, and M. Y. Jaffrin (2006). Treatment of dairy process waters modelled by diluted milk using dynamic nanofiltration with a rotating disk module. *Journal of Membrane Science 282*(1-2), 465–472. doi:10.1016/j.memsci.2006.06.005.

[30] Yorgun, M., I. A. Balcioglu, and O. Saygin (2008). Performance comparison of ultrafiltration, nanofiltration and reverse osmosis on whey treatment. *Desalination 229*(1), 204–216. doi:10.1016/j.desal.2007.09.008.

[31] Vourch, M., B. Balannec, B. Chaufer, and G. Dorange (2008). Treatment of dairy industry wastewater by reverse osmosis for water reuse. *Desalination 219*, 190–202. doi:10.1016/j.desal.2007.05.013.

[32] Frappart, M., M. Jaffrin, and L. H. Ding (2008). Reverse osmosis of diluted skim milk: Comparison of results obtained from vibratory and rotating disk modules. *Separation and Purification Technology 60*(3), 321–329. doi:10.1016/j.seppur.2007.09.007.

[33] Luo, J., L. Ding, Y. Wan, P. Paullier, and M. Y. Jaffrin (2010). Application of nf-rdm (nanofiltration rotating disk membrane) module under extreme hydraulic conditions for the treatment of dairy wastewater. *Chemical Engineering Journal 163*(3), 307–316. doi:10.1016/j.cej.2010.08.007.

[34] Luo, J. and L. Ding (2011). Influence of ph on treatment of dairy wastewater by nanofiltration using shear-enhanced filtration system. *Desalination 278*(1-3), 150–156. doi:10.1016/j.desal.2011.05.025.

[35] Luo, J., L. Ding, B. Qi, M. Y. Jaffrin, and Y. Wan (2011). A two-stage ultrafiltration and nanofiltration process for recycling dairy wastewater. *Bioresource Technology 102*(16), 7437–7442. doi:10.1016/j.biortech.2011.05.012.

[36] Gong, Y.-W., H.-X. Zhang, and X.-N. Cheng (2012). Treatment of dairy wastewater by two-stage membrane operation with ultrafiltration and nanofiltration. *Water Science and Technology 65*(5), 915–919. doi:10.2166/wst.2012.937.

[37] Hepsen, R. and Y. Kaya (2012). Optimization of membrane fouling using experimental design: an example from dairy wastewater treatment. *Industrial & Engineering Chemistry Research 51*, 16074–16084. doi:10.1021/ie302450s.

[38] Riera, F. A., A. Suárez, and C. Muro (2013). Nanofiltration of uht flash cooler condensates from a dairy factory: Characterisation and water reuse potential. *Desalination 309*, 52–63. doi:10.1016/j.desal.2012.09.016.

[39] Aydiner, C., S. Topcu, C. Tortop, F. Kuvvet, D. Ekinci, N. Dizge, and B. Keskinler (2013). A novel implementation of water recovery from whey: "forward-reverse osmosis" integrated membrane system. *Desalination and Water Treatment 51*(4-6), 786–799. doi:10.1080/19443994.2012.693713.

[40] Zmievskii, Y. G., I. I. Kirichuk, and V. G. Mironchuk (2014). Membrane treatment of wastewater obtained after the whey processing. *Journal of Water Chemistry and Technology 36*(6), 309–316. doi:Doi 10.3103/S1063455x14060095.

[41] Suárez, A., T. Fidalgo, M. A. Berdasco, and F. A. Riera (2014). Uht condensate recovery by reverse osmosis: A pilot-plant study. *Industrial and Engineering Chemistry Research 53*(39), 15237–15244. doi:10.1021/ie5026166.

[42] Suárez, A., T. Fidalgo, and F. A. Riera (2014). Recovery of dairy industry wastewaters by reverse osmosis. production of boiler water. *Separation and Purification Technology 133*, 204–211. doi:10.1016/j.seppur.2014.06.041.

[43] Suárez, A. and F. A. Riera (2015). Production of high-quality water by reverse osmosis of milk dairy condensates. *Journal of Industrial and Engineering Chemistry 21*, 1340–1349. doi:10.1016/j.jiec.2014.06.004.

[44] Andrade, L. H., F. D. S. Mendes, J. C. Espindola, and M. C. S. Amaral (2015). Reuse of dairy wastewater treated by membrane bioreactor and nanofiltration: Technical and economic feasibility. *Brazilian Journal of Chemical Engineering 32*(3), 735–747. doi:10.1590/0104-6632.20150323s00003133.

[45] Richelieu, M., U. Houlberg, and J. C. Nielsen (1997). Determination of alpha-acetolactic acid and volatile compounds by headspace gas chromatography. *Journal of Dairy Science 80*(9), 1918–1925. doi:10.3168/jds.S0022-0302(97)76132-0.

[46] Cortés-Francisco, N. and J. Caixach (2013). Molecular characterization of dissolved organic matter through a desalination process by high resolution mass spectrometry. *Environmental Science & Technology 47*(17), 9619–9627. doi:10.1021/es4000388.

[47] Fane, A. G., C. J. D. Fell, and A. G. Waters (1983). Ultrafiltration of protein solutions through partially permeable membranes — the effect of adsorption and solution environment. *Journal of Membrane Science 16*, 211–224. doi:10.1016/S0376-7388(00)81311-3.

[48] Husted, S., D. P. Persson, K. H. Laursen, T. H. Hansen, P. Pedas, M. Schiller, J. N. Hegelund, and J. K. Schjoerring (2011). Review: The role of atomic spectrometry in plant science. *Journal of Analytical Atomic Spectrometry 26*(1), 52–79. doi:10.1039/C0JA00058B.

[49] Cath, T. Y., A. E. Childress, and M. Elimelech (2006). Forward osmosis: Principles, applications, and recent developments. *Journal of Membrane Science 281*(1-2), 70–87. doi:10.1016/j.memsci.2006.05.048.

[50] Aydiner, C., U. Sen, S. Topcu, D. Ekinci, A. D. Altinay, D. Y. Koseoglu-Imer, and B. Keskinler (2014). Techno-economic viability of innovative membrane systems in water and mass recovery from dairy wastewater. *Journal of Membrane Science 458*, 66–75. doi:10.1016/j.memsci.2014.01.058.

[51] Tang, C., Z. Wang, I. Petrinić, A. G. Fane, and C. Hélix-Nielsen (2015). Biomimetic aquaporin membranes coming of age. *Desalination 368*, 89–105. doi:10.1016/j.desal.2015.04.026.

[52] Kalafatakis, S., S. Braekevelt, V. Carlsen, L. Lange, I. V. Skiadas, and H. N. Gavala (2017). On a novel strategy for water recovery and recirculation in biorefineries through application of forward osmosis membranes. *Chemical Engineering Journal 311*, 209–216. doi:10.1016/j.cej.2016.11.092.

[53] Adham, S., P. Gagliardo, D. Smith, D. Ross, K. Gramith, and R. Trussell (1998). Monitoring the integrity of reverse osmosis membranes. *Desalination 119*(1-3), 143–150. doi:10.1016/S0011-9164(98)00134-9.

[54] Kumar, M., S. Adham, and J. DeCarolis (2007). Reverse osmosis integrity monitoring. *Desalination 214*(1-3), 138–149. doi:10.1016/j.desal.2006.10.021.

[55] Pype, M. L., M. G. Lawrence, J. Keller, and W. Gernjak (2016). Reverse osmosis integrity monitoring in water reuse: The challenge to verify virus removal - a review. *Water Research 98*, 384–395. doi:10.1016/j.watres.2016.04.040.

[56] Henderson, R. K., A. Baker, K. R. Murphy, A. Hambly, R. M. Stuetz, and S. J. Khan (2009). Fluorescence as a potential monitoring tool for recycled water systems: A review. *Water Research 43*(4), 863–881. doi:10.1016/j.watres.2008.11.027.

[57] Singh, S., R. K. Henderson, A. Baker, R. M. Stuetz, and S. J. Khan (2012). Characterisation of reverse osmosis permeates from municipal recycled water systems using fluorescence spectroscopy: Implications for integrity monitoring. *Journal of Membrane Science 421-422*, 180–189. doi:10.1016/j.memsci.2012.07.008.

[58] Pype, M. L., D. Patureau, N. Wery, Y. Poussade, and W. Gernjak (2013). Monitoring reverse osmosis performance: Conductivity versus fluorescence excitation-emission matrix (eem). *Journal of Membrane Science 428*, 205–211. doi:10.1016/j.memsci.2012.10.027.

[59] Olesberg, J. T., M. A. Arnold, and M. J. Flanigan (2004). Online measurement of urea concentration in spent dialysate during hemodialysis. *Clinical Chemistry 50*, 175–181. doi:10.1373/clinchem.2003.025569.

[60] Hall, J. W. and A. Pollard (1993). Near infrared analysis of dried human serum for determining human serum urea concentration. *Journal of Near Infrared Spectroscopy 132*(1), 6. doi:10.1255/jnirs.15.

[61] Eddy, C. V., M. Flanigan, and M. A. Arnold (2003). Near-infrared spectroscopic measurement of urea in dialysate samples collected during hemodialysis treatments. *Applied Spectroscopy 57*(10), 1230–1235. doi:10.1366/000370203769699081.

[62] Cho, D. S., J. T. Olesberg, M. J. Flanigan, and M. A. Arnold (2008). On-line near-infrared spectrometer to monitor urea removal in real time during hemodialysis. *Applied Spectroscopy 62*(8), 866–872. doi:10.1366/000370208785284411.

[63] Siesler, H. W. (2007). Basic principles of near-infrared spectroscopy. In D. A. Burns and E. W. Ciurczak (Eds.), *Handbook of Near-Infrared Analysis* (3rd ed.)., Chapter 2. CRC Press.

[64] Krämer, N. and M. Sugiyama (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association 106*(494), 697–705.

[65] Currie, L. A. (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities. *Pure & Applied Chemistry 67*(10), 1699–1723. doi:10.1016/S0003-2670(99)00104-X.

[66] Lorber, A., K. Faber, and B. R. Kowalski (1997). Net analyte signal calculation in multivariate calibration. *Anal. Chem. 69*(8), 1620–1626. doi:10.1021/ac960862b.

[67] Bro, R. and C. M. Andersen (2003). Theory of net analyte signal vectors in inverse regression. *Journal of Chemometrics 17*(12), 646–652. doi:10.1002/cem.832.

[68] Kalivas, J. H. (2009). *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Chapter 3.01 Calibration Methodologies. Elsevier.

[69] Boqué, R., M. S. Larrechi, and F. X. Rius (1999). Multivariate detection limits with fixed probabilities of error. *Chemometrics and Intelligent Laboratory Systems 45*, 397–408. doi:10.1016/S0169-7439(98)00195-6.

[70] Allegrini, F. and A. C. Olivieri (2014). Iupac-consistent approach to the limit of detection in partial least-squares calibration. *Analytical Chemistry 86*, 7858–7866. doi:10.1021/ac501786u.

[71] Büning-Pfaue, H. (2003). Analysis of water in food by near infrared spectroscopy. *Food Chemistry 82*(1), 107–115. doi:10.1016/S0308-8146(02)00583-6.

[72] Wülfert, F., W. T. Kok, and A. K. Smilde (1998). Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Analytical Chemistry 70*(9), 1761–1767. doi:10.1021/ac9709920.

[73] Svendsen, C., T. Skov, and F. W. van den Berg (2015). Monitoring fermentation processes using in-process measurements of different orders. *Journal of Chemical Technology and Biotechnology 90*(2), 244–254. doi:10.1002/jctb.4483.

[74] Sanchez, E. and B. R. Kowalski (1988a). Tensorial calibration: I. first-order calibration. *Journal of Chemometrics 2*(4), 247–263. doi:10.1002/cem.1180020404.

[75] Sanchez, E. and B. R. Kowalski (1988b). Tensorial calibration: Ii. second-order calibration. *Journal of Chemometrics 2*, 265–580. doi:10.1002/cem.1180020405.

[76] Peinado, A. C., F. van den Berg, M. Blanco, and R. Bro (2006). Temperature-induced variation for NIR tensor-based calibration. *Chemometrics and Intelligent Laboratory Systems 83*(1), 75–82. doi:10.1016/j.chemolab.2006.01.006.

[77] Bro, R. (1997). Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems 38*(2), 149–171. doi:10.1016/S0169-7439(97)00032-4.

[78] Smilde, A., R. Bro, and P. Geladi (2004). *Multi-way Analysis with Applications in the Chemical Sciences*. John Wiley & Sons, Ltd.

[79] Bro, R. and H. A. L. Kiers (2003). A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics 17*(5), 274–286. doi:10.1002/cem.801.

[80] Buflovak falling film evaporators. `http://www.buflovak.com/products/evaporation-distillation-separation-solutions/falling-film-evaporator/`. accessed 2017-09-22.

[81] Atkins, P. and J. de Paula (Eds.) (2006). *Physical Chemistry for the Life Sciences*, Chapter 13 Optical Spectroscopy and Photobiology. Oxford University press.

[82] Sumner, L. W., A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W. M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, and M. R. Viant (2007). Proposed minimum reporting standards for chemical analysis: Chemical analysis working group (cawg) metabolomics standards initiative (msi). *Metabolomics 3*(3), 211–221. doi:10.1007/s11306-007-0082-2.

[83] Souza-Silva, É. A., R. Jiang, A. Rodríguez-Lafuente, E. Gionfriddo, and J. Pawliszyn (2015). A critical review of the state of the art of solid-phase microextraction of complex matrices i. environmental analysis. *Trends in Analytical Chemistry 71*, 224–235. doi:10.1016/j.trac.2015.04.016.

[84] Camino-Sánchez, F. J., R. Rodríguez-Gómez, A. Zafra-Gómez, A. Santos-Fandila, and J. L. Vílchez (2014). Stir bar sorptive extraction: Recent applications, limitations and future trends. *Talanta 130*, 388–399. doi:10.1016/j.talanta.2014.07.022.

[85] Prieto, A., O. Basauri, R. Rodil, A. Usobiaga, L. A. Fernández, N. Etxebarria, and O. Zuloaga (2010). Stir-bar sorptive extraction: A view on method optimisation, novel applications, limitations and potential solutions. *Journal of Chromatography A*, 2642–2666. doi:10.1016/j.chroma.2009.12.051.

[86] Gumbel, E. J. (1958). *Statistics of Extremes*. New York: Columbia University Press.

[87] Gomes, M. I. and A. Guillou (2015). Extreme value theory and statistics of univariate extremes: A review. *International Statistical Review 83*(2), 263–292. doi:10.1111/insr.12058.

[88] Abild, J., E. Y. Andersen, and D. Rosbjerg (1992). The climate of extreme winds at the great belt, denmark. *Journal of Wind Engineering and Industrial Aerodynamics 41*(1-3), 521–532. doi:10.1016/0167-6105(92)90458-M.

[89] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values* (1sted.). Springer Series in Statistics. London: Springer.

[90] Khalifa, M., F. Khan, and M. Haddara (2012). A methodology for calculating sample size to assess localized corrosion of process components. *Journal of Loss Prevention in the Process Industries 25*(1), 70–80. doi:10.1016/j.jlp.2011.07.004.

[91] Bhandari, J., F. Khan, R. Abbassi, V. Garaniya, and R. Ojeda (2015). Modelling of pitting corrosion in marine and offshore steel structures - a technical review. *Journal of Loss Prevention in the Process Industries 37*, 39–62. doi:10.1016/j.jlp.2015.06.008.

[92] Cook, N. J. (1982). Towards better estimation of extreme winds. *Journal of Wind Engineering and Industrial Aerodynamics 9*(3), 295–323. doi:10.1016/0167-6105(82)90021-6.

[93] Jensen, M. and N. Franck (1970). The climate of strong winds in denmark. Technical report, Danish Technical Press.

[94] Naess, A. and P. H. Clausen (2001). Combination of the peaks-over-threshold and bootstrapping methods for extreme value prediction. *Structural Safety 23*(4), 315–330. doi:10.1016/S0167-4730(02)00015-2.

[95] Wehrens, R., H. Putter, and L. M. C. Buydens (2000). The bootstrap: A tutorial. *Chemometrics and Intelligent Laboratory Systems 54*(1), 35–52. doi:10.1016/S0169-7439(00)00102-7.

[96] Nielsen, E. W. (1998). Long term use of a cheddar starter and development of phages with homology to its bacteria. *International Dairy Journal 8*, 1003–1009.

[97] Coffey, A. and R. P. Ross (2002). Bacteriophage-resistance systems in dairy starter strains: molecular analysis to application. *Antonie Van Leeuwenhoek 82*, 303–321.

[98] Kleppen, H. P., T. Bang, I. F. Nes, and H. Holo (2011). Bacteriophages in milk fermentations: Diversity fluctuations of normal and failed fermentations. *International Dairy Journal 21*(9), 592–600.

Appendix

Paper IV

Gas chromatography – mass spectrometry data processing made easy

Lea G. Johnsen, Peter B. Skou, Bekzod Khakimov and Rasmus Bro

Journal of Chromatography A, 2017, 57–64

Journal of Chromatography A, 1503 (2017) 57–64

Contents lists available at ScienceDirect

Journal of Chromatography A

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

Gas chromatography – mass spectrometry data processing made easy

Lea G. Johnsen[a,*], Peter B. Skou[b], Bekzod Khakimov[b], Rasmus Bro[b]

[a] MS-Omics, Birkehegnet 40, Ålsgårde, Denmark
[b] Copenhagen University, Thorvaldsensvej 40, Frederiksberg, Denmark

ARTICLE INFO

Article history:
Received 22 January 2017
Received in revised form 24 April 2017
Accepted 25 April 2017
Available online 27 April 2017

Keywords:
PARAFAC2
Chromatography
Data processing
Deconvolution
GC–MS

ABSTRACT

Evaluation of GC–MS data may be challenging due to the high complexity of data including overlapped, embedded, retention time shifted and low S/N ratio peaks. In this work, we demonstrate a new approach, PARAFAC2 based Deconvolution and Identification System (PARADISe), for processing raw GC–MS data. PARADISe is a computer platform independent freely available software incorporating a number of newly developed algorithms in a coherent framework. It offers a solution for analysts dealing with complex chromatographic data. It allows extraction of chemical/metabolite information directly from the raw data. Using PARADISe requires only few inputs from the analyst to process GC–MS data and subsequently converts raw netCDF data files into a compiled peak table. Furthermore, the method is generally robust towards minor variations in the input parameters. The method automatically performs peak identification based on deconvoluted mass spectra using integrated NIST search engine and generates an identification report. In this paper, we compare PARADISe with AMDIS and ChromaTOF in terms of peak quantification and show that PARADISe is more robust to user-defined settings and that these are easier (and much fewer) to set. PARADISe is based on non-proprietary scientifically evaluated approaches and we here show that PARADISe can handle more overlapping signals, lower signal-to-noise peaks and do so in a manner that requires only about an hours worth of work regardless of the number of samples. We also show that there are no non-detects in PARADISe, meaning that all compounds are detected in all samples.

1. Introduction

In chromatographic methods, such as gas or liquid chromatography coupled with mass spectrometry detectors, the goal is to identify compounds and compare their concentrations across and within samples. To achieve this goal, data processing must fulfil two criteria: (I) it must correctly determine the mass spectrum of the individual compounds for identification and; (II) it must accurately calculate the abundance of chromatographic peaks corresponding to those compounds in each sample. These two tasks are often challenging and time consuming mainly due to the co-elution of chromatographic peaks within a single chromatogram, as well as retention time (RT) shift of peaks across samples. These two challenges lead to mixed mass spectra and complicates compound identification and quantification. For these reasons processing of GC–MS data is challenging using currently available techniques that may perform inadequately both with respect to identification and quantification leading to compounds being wrongly interpreted or simply left undetected.

Most traditional vendor software quantifies compounds based on peak area or height using total ion count (TIC), base peak chromatogram (BPC) or from the extracted ion chromatogram (EIC) by selecting m/z value(s) typical for the given compound. These approaches are susceptible to co-eluting compounds since a contribution to the signal from other compounds is not adequately handled and may significantly affect both quantitative and qualitative results. Furthermore, it is challenging to estimate baseline contributions and this may also lead to errors in quantification. Most of currently applied approaches use simple subtraction of background from nearby baseline or a shoulder of a given peak of interest. Often this is not sufficient to handle overlapping and/or co-eluting peaks.

A more recent approach dealing with overlapping signals is to model the signals using e.g. Gaussian curves [1]. However, these models are not unique [2], instead, a number (actually infinitely many) of completely different sets of Gaussian peaks can model the data equally well. Hence, the solution becomes arbitrary. The development of the software package Automatic Mass spectral Deconvolution and Identification System (AMDIS) [3] was a big step towards resolving complex data. AMDIS automatically calculates the area of the deconvoluted component in terms of the area of the reconstructed total ion current (TIC) chromatogram. AMDIS is freely available standalone software, and is also implemented in

* Corresponding author.
E-mail addresses: lgj@msomics.com (L.G. Johnsen), peter.b.skou@food.ku.dk (P.B. Skou), bzo@food.ku.dk (B. Khakimov), rb@life.ku.dk (R. Bro).

http://dx.doi.org/10.1016/j.chroma.2017.04.052

commercial software like Masshunter (Agilent Technologies, USA). Another commercial software is ChromaTOF (LECO Inc., USA) that became a common tool to process GC–MS data based on a Time-Of-Flight (TOF) mass analyser. Like in AMDIS, ChromaTOF performs automatic deconvolution of peaks from each sample separately and compares the deconvoluted spectra against integrated libraries. Estimation of the peak area in ChromaTOF can either be based on the TIC, BPC, deconvoluted mass spectra or any *m*/*z* ion(s) that are defined by the user. ChromaTOF utilises a proprietary deconvolution technique, but it requires several input parameters, concerning noise level, peak width, retention time shift allowance and more, to be set by the user depending on the sample type and data quality. After peak detection, ChromaTOF can generate the final metabolite table by aligning peaks across samples based on user defined parameters such as RT shift window, noise level, spectral similarity and how often peaks are detected among investigated samples. Both AMDIS and ChromaTOF perform calculations on each sample independently of the other samples.

A completely different approach for handling co-elution and retention time shifts, is to use the so-called PARAllel FACtor analysis2 (PARAFAC2) model [2,4]. PARAFAC2 is able to deconvolute co-eluted, retention time shifted and low signal-to-noise (S/N) ratio chromatographic peaks for all investigated samples in a given retention time region simultaneously [2]. In contrast to other methods, the PARAFAC2 approach only requires a single parameter to be set by the user prior to achieving sufficient data processing for the given retention time region of the chromatogram. This parameter is the number of factors (or real chemical compounds) in the investigated region of the chromatogram. There are simple methods for determining this number as will be explained later. PARAFAC2 modelling allows extraction of the pure spectra of co-eluting compounds as well as it simultaneously computes their peak areas (relative concentrations). The compounds are quantified using the entire pure spectrum and retention time region corresponding to a specific peak. It has previously been shown that PARAFAC2 is superior to commercial solutions [5,6]. However, current implementations of PARAFAC2 are not accessible for non-mathematical users and requires extensive coding for efficient use. Here, we develop an integrated approach called PARAFAC2 based Deconvolution and Identification System (PARADISe), which combines workflow from raw data inspection to metabolite (relative) quantification and identification in a graphical user interface (GUI). Within the PARADISe approach, we included tools required in all steps of the GC–MS data processing; 1) data visualization, 2) division of data into retention time intervals, 3) PARAFAC2 based deconvolution of peaks, 4) validation and extraction of deconvoluted peaks, 5) identification of compounds from raw as well as deconvoluted mass spectra using NIST search engine and NIST mass spectra library and/or any other libraries in NIST format, 6) generation of the final metabolite table. In the following sections, several examples are provided illustrating the power and limits of PARADISe.

2. Materials and methods

2.1. Preparation of a standard mixture sample

Ten chemical compounds including valine, alanine, serine, threonine, *gamma*-aminobutyric acid (GABA), ascorbic acid, fumaric acid, citric acid, gallic acid and *p*-hydroxyphenylacetic acid were used to prepare a standard mixture sample. Compounds were purchased from Sigma-Aldrich (Sigma-Aldrich Denmark A/S, DK) at the highest available purity. The standard mixture sample was prepared by mixing equal volumes of 20.0 mM solutions of compounds in milliQ water. Thus, in the final standard mixture sample the concentration of each compound was 2.0 mM, which was used for preparation of ten different dilution series samples where concentration of each compound ranged from 0.05 to 0.6 mM.

2.2. GC–MS analysis of standard mixture samples

Prior to GC–MS analysis 30 μL of each dilution series samples were dried using ScanVac (Labogene, DK) at 40 °C inside 150 μL glass inserts, sealed with air tight magnetic lids into GC–MS vials and derivatized by addition of 30 μL trimethylsilyl cyanide (TMSCN) [7]. All steps involving sample derivatization and injection were automated using a Dual-Rail MultiPurpose Sampler (MPS) (Gerstel, GmbH & Co. KG, DE). Following reagent addition, the sample was transferred into the agitator of the MPS and incubated at 40 °C for 40 min at 750 rpm. This procedure ensures precise derivatization time and reproducible sample injection. Immediately after derivatization, 1 μL of the derivatized sample was injected into a cooled injection system (CIS4, Gerstel, GmbH & Co. KG, DE) port in splitless mode. The septum purge flow and purge flow to split vent at 2.5 min after injection were set to 25 and 15 mL min^{-1}, respectively. Initial temperature of the CIS port was 40 °C, and heated at 12 °C s^{-1} to 320 °C (after 30 s of equilibrium time), where it was kept for 5 min. After heating, the CIS port was gradually cooled to 250 °C at 5 °C s^{-1}, and this temperature was kept constant during the run. A GC–MS consisted of an Agilent 7890 B gas chromatograph (GC) and a high-throughput Pegasus GC-TOF-MS mass spectrometer (LECO Inc. USA). More details of GC oven and cooled injection system (CIS4) condition were the same as previously described [7]. Mass spectra were recorded in the *m*/*z* range of 45–600 with a scanning frequency of ten scans sec^{-1}, and the MS detector and ion source were switched off during the first 4.5 min of solvent delay time. The transfer line and ion source temperature were set to 280 °C and 250 °C, respectively. The mass spectrometer was tuned according to manufacturer's recommendation using perfluorotributylamine (PFTBA). The MPS and GC–MS was controlled using vendor software Maestro (Gerstel, GmbH & Co. KG, DE) and ChromaTOF (LECO Inc., USA). Samples were randomised prior to derivatization and GC–MS analysis, and a blank sample containing only derivatization reagent, and an alkane mixture standard (all even C10-C40 alkanes at 50 mg L^{-1} in hexane) were analysed at least between five real samples prior to monitor GC–MS performance.

2.3. Analysis of complex samples

The dataset investigated in this study consisted of 69 samples including blank samples and pooled quality control samples. The complex samples are media samples obtained from fermentation of CHO cells in complex media, the cells are removed by filtration and the spent media is kept on −20 °C until the time of derivatization. Prior to the analysis, the samples were derivatized using a procedure based on the protocol described by Smart et al. [8]. All samples were analysed in a randomised order. A 6890N GC in conjunction with a 5975 B quadrupole mass spectrometer (Agilent Technologies, USA) were used to analyse the samples. The system was controlled by ChemStation (Agilent Technologies, USA).

3. Theory

PARADISe is based on PARAFAC2 modelling, which allows simultaneous deconvolution of pure mass spectra of peaks and integration of areas of deconvoluted peaks for all samples. Resolved peaks are identified using their deconvoluted pure mass spectra and the final peak table is generated. Thus, PARADISe is based on five major steps:

1. Define intervals
2. Resolve compounds
3. Validate models
4. Identify compounds
5. Create peak table

PARADISe, integrates all these as outlined below.

Intervals are selected manually through an interactive TIC plot in such a way that approximate baseline-resolved intervals, with preferably less than six peaks, are obtained. As will be illustrated later, the specific definition of the intervals is not critical (within reason). Having defined each interval, the PARAFAC2 model can resolve the underlying and possibly overlapping compounds in each of these intervals. For each interval, a separate PARAFAC2 model is built. To do so, the number of chemical compounds (including baseline) must be defined for the specific PARAFAC2 model. PARADISe will by default calculate models with one to eight components, and it is the user that must decide which of the models to use. Automated methods exist for determining the number of components [5] but in PARADISe, the user has to do this. Normally, the number of components is set to the highest number that still maintains a sufficiently high core consistency (above 50%). Visualizations of the models can be used for intervals that may pose special problems to further guide the user but this is mostly not critical. Once the model for a given interval is determined, compounds of interest can be tagged (e.g. compounds that are not baseline or tails from peaks surrounding the interval) and only these compounds will be included in the final report. For a more thorough description of the theory behind PARAFAC2 the reader is referred to the supplementary material.

The PARAFAC2 model of each compound provides the relative concentration (peak area) directly and users can evaluate elution profiles of deconvoluted peaks. Identification is also a crucial part of the chromatographic analysis, and PARADISe enables the user to make library lookups of both mass spectra from raw data and PARAFAC2 deconvoluted mass spectra (pure compound spectra). The lookup is performed by exporting relevant spectra to the NIST MSsearch, which therefore must be installed prior to use the library lookup function. The user can then perform the evaluation of any library hits directly in the MSsearch software.

PARADISe is built around two main interfaces; one, which is used for inspection of raw data and creation of intervals, and one, which is used to visualize and validate models prior to select deconvoluted peaks and to create a final report. The software is compiled via Matlab and is thus platform independent and can work without NIST software. However, using the PARADISe without the NIST software eliminates the possibility of performing library searches of mass spectra. An overview of the full workflow is illustrated in Fig. 1.

Two formats of raw data can currently be imported; either data in the cdf format for mass spectrometry, or for users who are familiar with Matlab, data can be imported from the Matlab format.

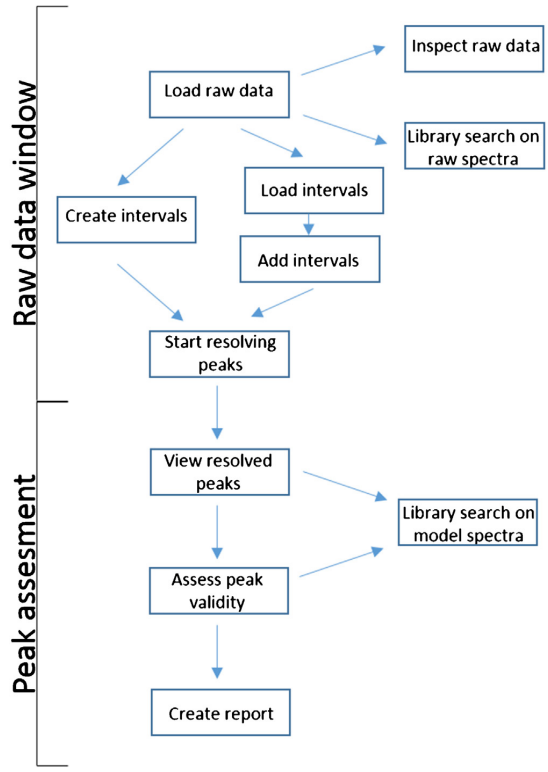

Fig. 1. Flowchart illustrating the workflow in PARADISe; from loading of raw data to generation of the final report with relative concentrations of detected compounds.

4. Results

In the following we will illustrate the capabilities of PARADISe through a number of small examples, each aimed at different typical challenges encountered in chromatographic data analysis.

4.1. Quantification

Quantification is an important part of data analysis. To illustrate the capabilities of PARADISe concerning quantitative determination of compounds, a dilution series of the standard mixture sample were analysed. The obtained data was processed using ChromaTOF, AMDIS and PARADISe (Fig. 2 and Fig. S2). All three software packages performed equally well when the S/N ratio of peaks was high. However, for the lower S/N ratio peaks, AMDIS and ChromaTOF results were sensitive to the settings of the user-defined parameters, while PARADISe performance was more consistent regardless of S/N ratio of peaks.

4.2. Co-elution

To demonstrate application of PARADISe to complex GC–MS profiles, a data set obtained from GC –MS analysis of spent media from cell cultures grown in complex media was investigated. One of the huge advantages of using PARADISe is its ability to deconvolute overlapping peaks. An example of the deconvolution power is illustrated in Fig. 3. The TIC of this data interval shows one peak, one baseline and one tail from a neighbouring peak. Upon inspection of the data using PARADISe, it becomes apparent that the interval is covering not one but three peaks and the interval is therefore best described with a five-component PARAFAC2 model: one component describing baseline, one the tail and one for each of the three peaks, respectively (see Fig. 3). Inspection of characteristic m/z ions (m/z 127, 216, and 130) of the deconvoluted peaks shows that the three peaks can be recognised from the corresponding extracted ion chromatograms (bottom plots in Fig. 3). It is worth to mention here that PARADISe allows such a deconvolution and provides pure spectra of deconvoluted peaks for even more complex chromatographic data intervals, without any user defined settings, besides the number of components.

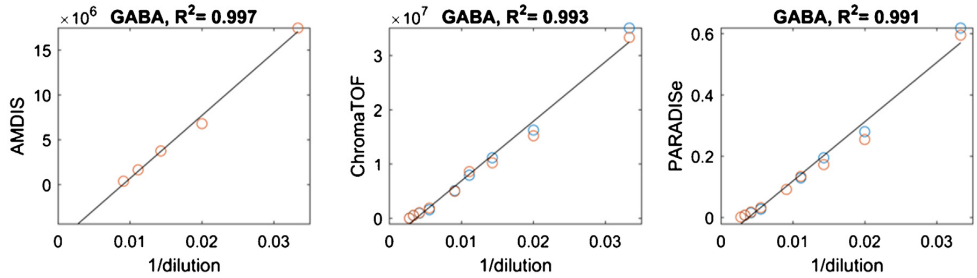

Fig. 2. Illustration of obtained relative concentrations from AMDIS, ChromaTOF and PARADISefrom dilution series analysis of GABA (red and blue correspond to replicates). GABA was not detected by AMDIS in the most diluted samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

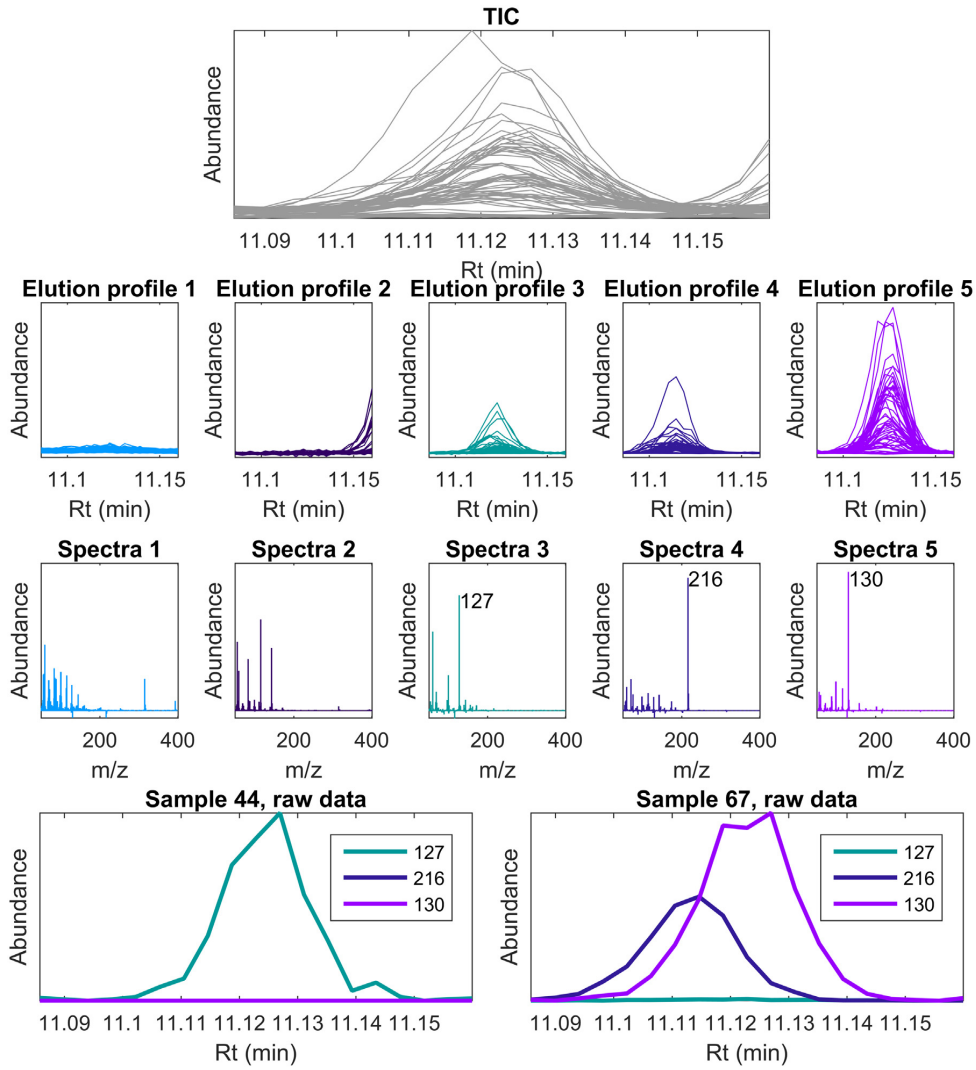

Fig. 3. Top: TIC of the interval, row 2: obtained elution profiles from a five-component model, row 3: model spectra obtained from the five-component model. Row 4: EIC of characteristic masses from the model (extracted from raw data).

Table 1
An overview of the data processing steps that require user defined parameters in three different GC–MS data processing software, AMDIS, ChromaTOF, and PARADISe. The number of hourglass indicates how many parameters must be set by the user in the given step of data proccessing, (–) indicates that this step is not performed by the software, and empty cells illustrate steps that do not require any parameters to be set by users for the given software.

Data proccessing steps that require parameters to be set by the user	Software		
	AMDIS	ChromaTOF	PARADISe
1) Define RT intervals for processing	–	–	⧗
2) Deconvolution	⧗⧗⧗⧗	⧗⧗⧗	
3) Peak filtering and removing baseline	⧗⧗⧗⧗⧗	⧗⧗⧗⧗	
4) Mass spectrometer dependent parameters	⧗⧗⧗		
5) Processed data validation			⧗
6) Alignment of peaks across samples	–	⧗⧗⧗⧗	

4.3. Low signal-to-noise

In contrast to other approaches, PARADISe is not so sensitive to the S/N ratio of peaks and is able to deconvolute extremely small peaks directly from the raw data (Figs. 4 and 5). In Fig. 4, the PARADISe results reveal that the investigated noisy interval actually contains two overlapping peaks with a very low S/N ratio. Inspection of characteristic m/z values in the raw data confirms that, within the given interval, two compounds are eluting with different mass spectra. Subsequently, a four-component PARAFAC2 model deconvoluted two peaks corresponding to two chemicals plus two components reflecting the background.

The second example (Fig. 5) shows how well the mass spectra from a low S/N ratio peak is modelled using PARADISe. Despite extreme low S/N ratio of this peak, its deconvoluted mass spectrum allowed identification using the NIST mass spectral library, found as dimethyl malonic acid. The identity of this compound was validated with an authentic standard, which was found to have the same retention time and mass spectrum.

4.4. Baseline

Baseline contributions present in a raw GC–MS data heavily influence both peak identification and quantification, thus it is important that data processing techniques can remove baseline contributions. In the model illustrated in Fig. 4 two different baselines are present and shows that it is possible to automatically remove these artefacts using PARADISe. It is often seen that the baseline is modelled using more than one PARAFAC2 factor, because the background is often a mixture of several contributions (e.g. column bleed, derivatization reagent, mobile phase, or electronic noise) All models presented in this paper illustrate how the PARADISe approach removes baseline contributions as separate PARAFAC2 components from eluting compounds eliminating any need for raw data pre-treatment.

4.5. Retention time drift

In the examples illustrated throughout this paper, different degrees of shift in RT are present (see Figs. 3–5). In all cases, PARADISe handles the drift without any prior assumptions about maximum allowed shift. PARADISe is also able to correctly determine peaks that have severe RT shifts across samples that sometimes result in complete cross RT shifts with nearby eluting peaks as well as with co-eluting peaks. This is only possible due to the unique mass spectrum of each compound and flexibility of deconvolution engine, PARAFAC2. However, in order to correctly determine all peaks present in a given chromatographic data interval, the width of the interval must be wide enough to cover RT shifts.

4.6. Limitations

There are two major cases when PARADISe fails to deconvolute GC–MS peaks: 1) when a GC–MS data interval contains two or more peaks with identical mass spectra, 2) when a GC–MS data interval contains two or more peaks that co-vary completely in their concentrations. In both cases PARADISe will find those co-varying peaks as a single compound. In the example illustrated in Fig. 6, two of the four peaks are lumped into one common component (Elution profile 4). This happens regardless of how many PARAFAC2 components are included in the model. Inspection of the raw data reveals that the two peaks have identical mass spectra (Top two rows, right in Fig. 6). It is a premise of PARAFAC2 that each chemical compound in a given interval must have at least slightly different spectral signature. Hence, when two compounds have identical spectra as here, they cannot be separated in a PARAFAC2 model. The only alternatives then are either 1) to split the data in between the two peaks, or 2) try to separate the peaks by other means (chemically or mathematically).

One can also choose either to exclude the compound from the final data set, or to use it, bearing in mind, that the reported concentration profile/spectra will be a combination of both peaks. Working within smaller retention time intervals minimizes the risk of modelling problems if different peaks co-vary across samples.

5. Discussion

PARADISe excels in simplicity because only little input is needed from the user to obtain valid models of the compounds and the inputs typically have a feasible range of settings so that the exact choice is not critical. The data must be split into retention time

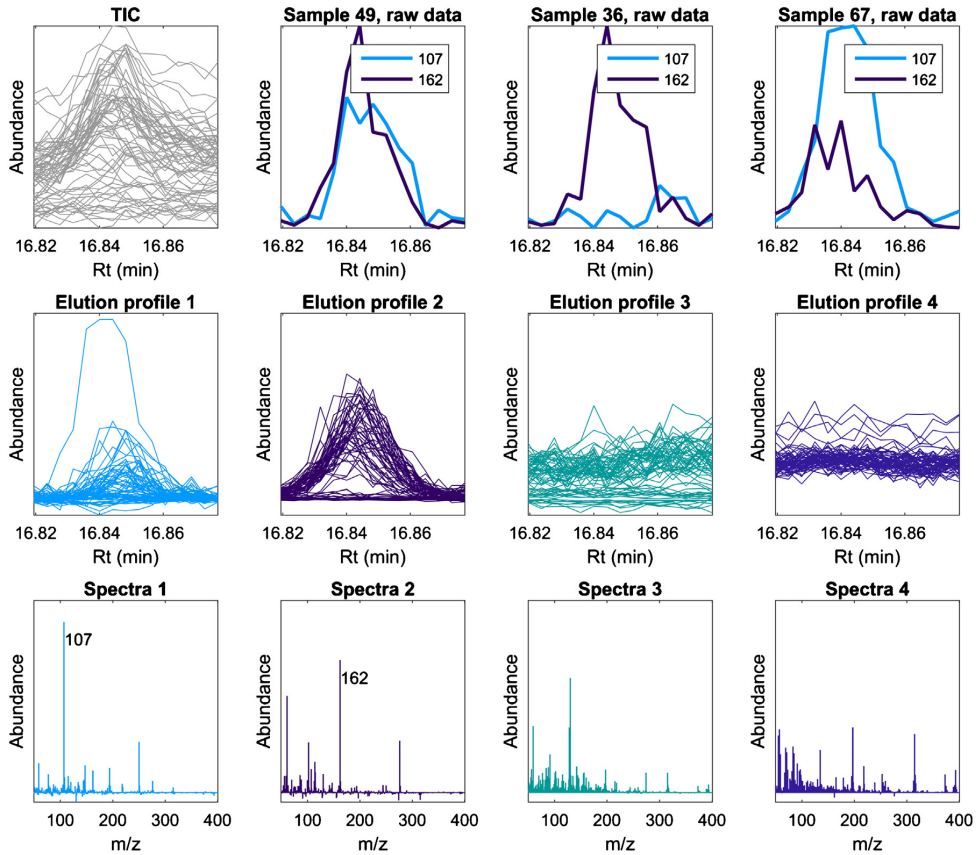

Fig. 4. Top left: TIC from raw data. Top right: EIC from raw data of selected *m*/*z*. Middle: elution profiles obtained from a four-component model. Elution profile 3 and 4 represent baseline. Bottom: spectra obtained from a four-component model.

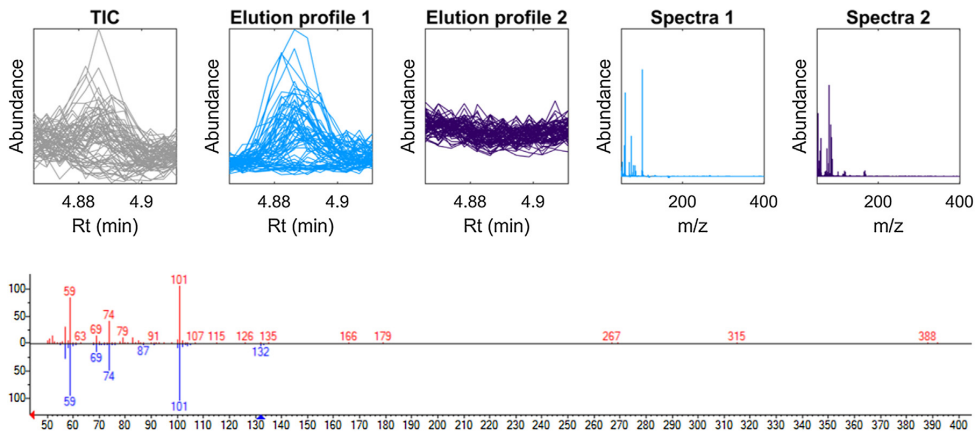

Fig. 5. Top: TIC from raw data, Elution profiles and spectra obtained from a two-component model. Bottom: comparison between the model spectra 1 and the NIST library spectra of dimethyl malonic acid. Profile 2 is representing baseline.

intervals with approximate baseline separation. The interval borders should be determined in a reasonable manner, meaning that the peaks of interest should be included in the interval without cutting off any tailing or fronting. Even tails from peaks adjacent to the intervals, as shown in Fig. S1, does not pose a problem. Further, as few compounds as possible should be included when selecting

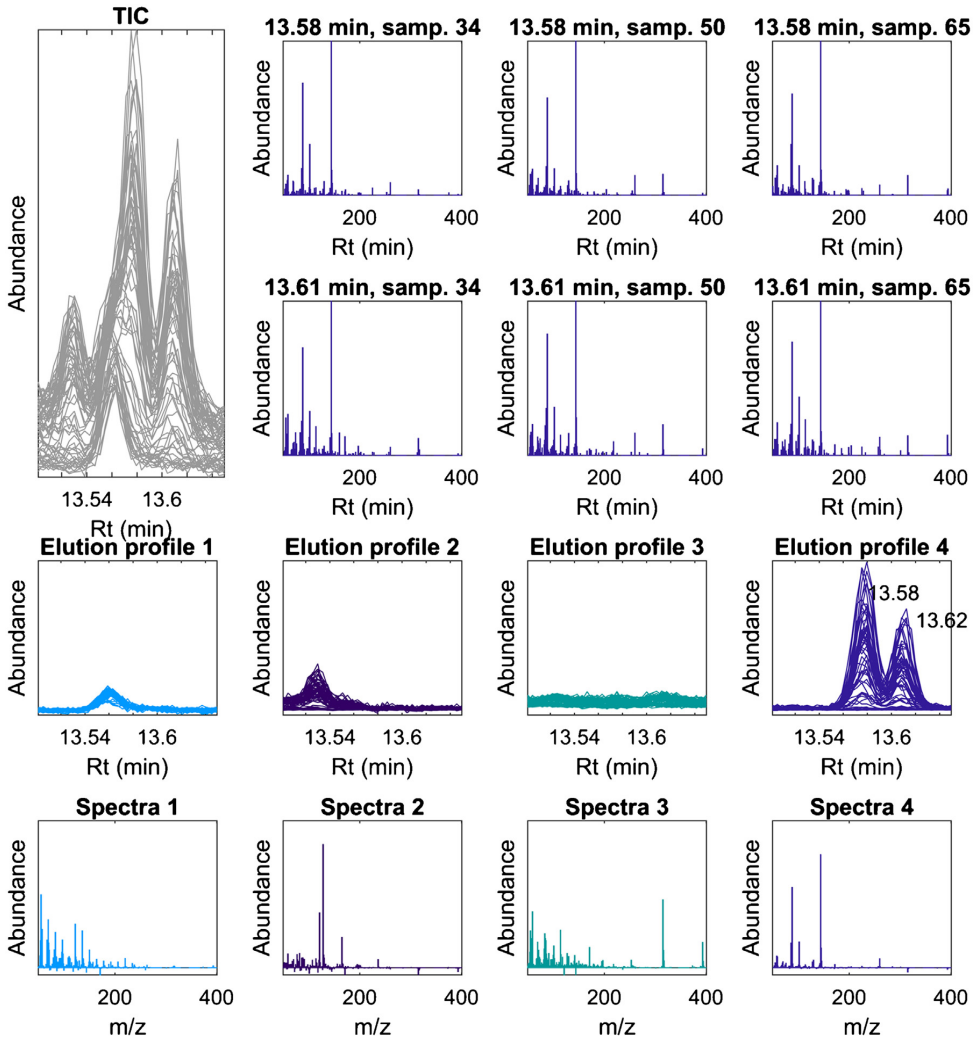

Fig. 6. Top left: TIC from raw data. Top 2 rows, right: spectra obtained from three different samples of the two peaks eluting at 13.58 and 13.62 min, respectively. Row 3: elution profiles obtained from a four-component model. Bottom row: spectra obtained from a four-component model.

intervals. Selecting a simpler (fewer compounds) interval reduces computation time and prevents small errors accumulating in more complicated models with many compounds.

Unlike some tools for processing of GC–MS data, the same model describes all samples when using PARADISe. This means that if a model is accepted as valid, all samples are well described in that particular interval and the developed method can routinely be applied to new samples without any user interaction.

An added benefit from using PARADISe is that there will not be any non-detects. In many methods, the user must specify parameters used to define a peak (e.g. peak width, signal to noise levels etc.). This means that if a peak does not match these criteria they will appear as "not detected". In most cases this will be due to a peak being lower than the limit of detection. These missing values will cause problems if the data is to be used in either classical statistics [9] or multivariate statistics. In more severe cases, a peak may actually be present but not fulfilling the initially set parameters. If the user does not recognize this, it will most likely be wrongly interpreted. In PARADISe there are no assumptions made about peak shape, signal-to-noise ratio or expected retention time shifts. When peaks are deconvoluted there will always be an estimate of the concentration (also in cases with signal being lower than the limit of detection), and the problems with missing values are therefore not an issue. In essence, the problem of non-detects is moved to the subsequent data analysis. All peaks are quantified and the possible decision of where to set the limit of detection can be decided after the quantification has been performed.

PARADISe cannot process one sample at a time but requires several samples prior to processing any dataset. It is not enough to analyse the same sample several times or to make dilutions of the samples and analyse these. If one wants to use PARADISe at least five samples with independent variations must be included in the sample set and preferably more.

To be able to compare the user-friendliness of the software AMDIS, ChromaTOF and PARADISe, we divide the workflow into five parts below for easier comparison:

1) Define RT intervals for processing. Division of the chromatographic data into smaller RT intervals is needed for reducing complexity when processing data using PARADISe prior to obtain reliable deconvolution.
2) Deconvolution. The deconvolution step in AMDIS and ChromaTOF requires parameters such as peak width, resolution, sensitivity, and shape to be set by users. The number of components must be determined in PARADISe.
3) Peak filtering and removing baseline. The peak filtering step requires parameters like S/N ratio, mass threshold, baseline offset, minimum abundance in AMDIS and ChromTOF.
4) Mass spectrometer dependent parameters. Mass spectrometer dependent parameters such as m/z range, scan direction, instrument type, file format, threshold are also crucial when using AMDIS.
5) Alignment of peaks across samples. Several parameters such as maximum allowed RT shift, spectral similarity, detection frequency (e.g., a peak must be present at least in 50% of samples) are required in ChromaTOF when aligning peaks across samples prior to a final metabolite table.

In Table 1 a summary is given, indicating how many parameters the user needs to set for each step.

PARADISe can be used for targeted analysis, where only the target compounds are processed, as well as untargeted analysis. In cases with routine targeted high-throughput GC–MS methods, interval-files can be predefined and reused. However, it is important to stress that the user should still inspect the raw data before processing the data. This is, in fact, underestimated in many data processing software packages, but we strongly advice data inspection prior to use PARADISe.

6. Conclusions

We have demonstrated a new approach called PARAllel factor analysis 2 based Deconvolution and Identification System (PARADISe), integrating multi-way modelling for processing of raw GC–MS data from several samples simultaneously. PARADISe combines entire workflow from raw data inspection to peak deconvolution and metabolite identification in a graphical user interface. It allows handling very complex situations with severe co-elution even with resolution close to zero. With PARADISe, a single standalone platform is presented covering the entire workflow from inspecting raw data to identification, including deconvolution of peaks across all samples simultaneously, determination of relative concentrations and compilation of a compound table. The ability to export mass spectra, deconvoluted (pure) as well as raw, to spectral databases can save large amounts of time and will increase the hit quality.

Acknowledgements

We would like to thank Dr. W. Gary Mallard from the National Institute of Standards and Technology (NIST) from sharing insights into how to optimally use the AMDIS software. It was a great help and we are very thankful for your time.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chroma.2017.04.052.

References

[1] M.L. Phillips, R.L. White, Dependence of chromatogram peak areas obtained by curve-fitting on the choice of peak shape function, J. Chromatogr. Sci. 35 (1997) 75–81.
[2] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605, http://dx.doi.org/10.1021/cr900394n.
[3] S.E. Stein, An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data, J. Am. Soc. Mass Spectrom. 10 (1999) 770–781, http://dx.doi.org/10.1016/S1044-0305(99)00047-1.
[4] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2—part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309, http://dx.doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295:AID-CEM547>3.0.CO;2-Y.
[5] L.G. Johnsen, J.M. Amigo, T. Skov, R. Bro, Automated resolution of overlapping peaks in chromatographic data, J. Chemom. 28 (2014) 71–82, http://dx.doi.org/10.1002/cem.2575.
[6] T. Skov, R. Bro, Solving fundamental problems in chromatographic analysis, Anal. Bioanal. Chem. 390 (2007) 281–285, http://dx.doi.org/10.1007/s00216-007-1618-z.
[7] B. Khakimov, R.J. Mongi, K.M. Sørensen, B.K. Ndabikunze, B.E. Chove, S.B. Engelsen, A comprehensive and comparative GC–MS metabolomics study of non-volatiles in Tanzanian grown mango, pineapple, jackfruit, baobab and tamarind fruits, Food Chem. 213 (2016) 691–699, http://dx.doi.org/10.1016/j.foodchem.2016.07.005.
[8] K.F. Smart, R.B.M. Aggio, J.R. Van Houtte, S.G. Villas-Bôas, Analytical platform for metabolome analysis of microbial cells using methyl chloroformate derivatization followed by gas chromatography-mass spectrometry, Nat. Protoc. 5 (2010) 1709–1729, http://dx.doi.org/10.1038/nprot.2010.108.
[9] D.R. Helsel, Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it, Chemosphere 65 (2006) 2434–2439.

DEPARTMENT OF FOOD SCIENCE
FACULTY OF SCIENCE · UNIVERSITY OF COPENHAGEN
PHD THESIS 2017 · ISBN 978-87-7209-080-1

PETER BÆK SKOU

Process-water characterisation and quality monitoring in the dairy industry
– moving towards replacing potable water

Food processing requires large amounts of water to the extent that even in Denmark local scarcity is challenging the food production industry. In the dairy industry process-water – water reclaimed from existing process streams such as Cleaning-In-Place or cheese whey processing for protein recovery – is an obvious candidate to replace potable water while in-turn minimizing wastewater discharge. In this work the process-water stems from equipment already present in the dairy industry, namely membrane filtration permeate as well as evaporator condensate. To ensure safe-for-use, high quality process-water at all times key quality attributes must be identified and (ideally) monitored continuously in the future.

This thesis deals with the chemical characterization of selected process-water streams, and identifies and tests relevant measurement techniques. Chemical characterization was performed using analytical techniques developed in the field of metabolomics. The first investigations focused on membrane permeate, and led to the identification of urea as the main organic compound, next to low levels of other organic compounds. Near infrared spectroscopy as potential on-line analytical method was tested to monitor the variation. Evaporator condensate was also investigated, but no conclusive identification could be made. Finally, extreme value theory – a statistical tool set – was applied to demonstrate a new direction in monitoring and characterizing of process dynamics in the dairy and food industry.

SL grafik