Model-based Monitoring and Optimization of a Bio-based Process

André Fernandes Caroço, Ricardo

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
Doctor of Philosophy

DTU Chemical Engineering
Department of Chemical and Biochemical Engineering

Model-based Monitoring and Optimization of a Bio-based Process

Ricardo Fernandes Caroço

Kongens Lyngby 2019

PROSYS - Process and Systems Engineering Center
Department of Chemical and Biochemical Engineering
Technical University of Denmark

Søltofts Plads
Building 229
2800 Kongens Lyngby, Denmark
Phone +45 4525 2800
kt@kt.dtu.dk
http://www.kt.dtu.dk/

Abstract

The present thesis is concerned with establishing a scientific-based operational strategy for operations that utilize biological raw material. Bio-based processes deal with the unavoidable natural disturbances that entail the processing of biological feedstock. The final properties of a product are invariably linked with the initial properties of the raw material. Nowadays large segments of the industry operate in a heuristic recipe-driven way, dependent on rule-of-thumb experience which too often leads to batch-to-batch discrepancies. To this end, the thesis deals with the development of tools necessary for the implementation of a flexible operational strategy.

The tools are developed to fit the rationale of incorporating critical material attributes with a desired product quality target to acquire optimized process conditions. The optimization provides the conditions at which the process should run, for the raw material with the assessed critical material attributes, to achieve the desired quality. Furthermore, there is a need to assess whether both the defined conditions are ideal for the production system, and our predictions of product end quality are correct. Thus, a predictor correction through the incorporation of in-operation measurements is a necessary component for continuous process verification and improvement. An industrial case study of pectin production, focused on the batch extraction from citrus fruit peels, is developed in collaboration with CP Kelco.

Citrus peels are analysed to demonstrate the conceptual and performance differences of distinct quality assessment approaches. The analysis demonstrates the advantage of characterization through multivariate data analysis coupled with a complementary spectroscopic technique, near-infrared spectroscopy. The quantitative comparative analysis of three different approaches, discriminant classification based on expert-knowledge, unsupervised classification, and spectroscopic correlation with reference physicochemical variables, is performed in the same dataset context.

A mathematical model developed by Andersen et al. (2017) is considered for optimization of the process, taking into account raw material quality uncertainty. Before application, the model is evaluated through local sensitivity analysis. The impact of raw material uncertainty was further assessed through an uncertainty analysis, which quantified the variability of model predictions for three different types of fruit. The model provides a good agreement with the general depiction of the extraction phenomena. The critical operating parameters, i.e., temperature, pH and batch time, are then optimized in a deterministic manner to maximize the final pectin concentration while satisfying given requirements. A robust (worst-case) optimization is examined

to design the optimal operational conditions in consideration of the inherent uncertainty of feedstock and the desired product quality.

The model prediction corrections are made through the use of the continuous-discrete Kalman filter algorithm. A systematic approach to constructing this predictor with a desired performance is presented. Discrepancies between the measured outputs and the filter model are observed. These differences are in the initial state guess and the considered model parameters. Implementation corrections are proposed to cope with these challenges and are evaluated for production scale data.

Ultimately, the combination of these tools is showcased for a particular peel, and the impact of the proposed operational strategy is assessed.

Resumé

Den forelæggende afhandling omhandler etableringen af en evidensbaseret fremgangsmåde for enhedsoperationer med råmateriale fra biologisk ophav. Bio-baserede processer skal kunne håndterer de uundgåelige naturlige variationer, der fremkommer ved forarbejdning af biologisk råmateriale. De endelige egenskaber af et produkt er direkte forbundet med råmaterialets oprindelige egenskaber. I dag opererer store segmenter af Bio-branchen på en heuristisk, opskriftdreven måde, afhængig af tommelfingerregler og mavefornemmelser, som for ofte fører til afvigelser fra batch til batch. Med det i mende, handler den forelæggende afhandling om udviklingen af de nødvendige redskaber til gennemførelse af en fleksibel operationel strategi.

Værktøjerne er udviklet med henblik på at inkorporere kritiske materialeegenskaber med et ønsket produktkvalitetsmål for at opnå optimerede procesforhold. Optimeringen specificerer de betingelser, hvorved processen skal afvikles, for at opnå den ønskede kvalitet, med de kritiske materialeegenskaber for råmaterialet i betragtning. Derudover er der behov for at vurdere, om de definerede betingelser er ideelle for produktionssystemet, og hvorvidt forudsigelser af produktets slutkvalitet er korrekte. Således er det nødvendigt at inkorporere en metodisk korrektion af målinger i drift til kontinuerlig procesverifikation og forbedring. Et industrielt case studie af pektin produktion, fokuseret på batch pektin udvinding af skralder fra citrusfrugt, er udviklet i samarbejde med CP Kelco.

Citrusskræller analyseres for at demonstrere konceptuelle og præstationsforskelle i forskellige kvalitetsvurderingsmetoder. Analysen demonstrerer fordelen ved karakterisering gennem multivariat dataanalyse koblet med en komplementær spektroskopisk teknik, nær-infrarød spektroskopi. Den kvantitative sammenlignende analyse af tre forskellige fremgangsmåder, diskriminant klassificering baseret på ekspertviden, ukontrolleret klassificering og spektroskopisk korrelation med reference til fysisk-kemiske variabler udføres i samme kontekst.

En matematisk model udviklet af Andersen et al. (2017) betragtes for at optimere processen ved at tage hensyn til usikkerheden om råmaterialets kvalitet. Før anvendelse, evalueres modellen ved hjælp af lokal sensitivitetsanalyse. Virkningen af råstofusikkerhed blev yderligere vurderet gennem en usikkerhedsanalyse, som kvantificerede variabiliteten af modelforudsigelser for tre forskellige typer frugter. Modellen stemte godt overens med den generelle skildring af ekstraktionsfænomenerne. De kritiske driftsparametre, dvs. temperatur, pH og batch tid optimeres derefter på baggrund af et deterministisk approach for at maksimere den endelige pektinkon-

centration samtidigt med at givne kvalitetskrav overholdes. En robust (worst case) optimering undersøges for at designe de optimale driftsbetingelser under hensyntagen til kvalitetsusikkerheden af råmaterialet og den ønskede produktkvalitet.

Korrektion of Modelforudsigelserne foretages ved brug af den kontinuerlige diskrete Kalman filter algoritme. En systematisk tilgang til konstruktion af denne prognosedanner med en ønsket nøjagtighed er præsenteret. Forskelle mellem de målte outputs og filtermodellen observeres. Disse forskelle er i det oprindelige tilstandsgæt og de betragtede modelparametre. korrektioner foreslås at klare disse udfordringer for Implementering og evalueres for produktionsskala data.

Til sidst er kombinationen af disse værktøjer fremvist for en given citrusskræl, og virkningen af den foreslåede operationelle strategi vurderes.

Preface

This thesis was prepared at the Department of Chemical and Biochemical Engineering at the Technical University of Denmark in partial fulfillment of the requirements for acquiring a Doctor of Philosophy degree in Chemical and Biochemical Engineering. The work was developed in the period from October 2015 until February 2019 and was supervised by Associate Professor Jakob K. Huusom (DTU Chemical Engineering) and Associate Professor Jens Abildskov (DTU Chemical Engineering) and Dr. Paloma Santacoloma (CP Kelco Aps.) as co-supervisors. The work has been performed within the BIOPRO2 strategic research consortium and received financial support from Innovation Fund Denmark.

Acknowledgments

Firstly, my acknowledgment to my supervisors efforts and valuable contribution. Thank you Jakob for the scientific and project management discussions, which enabled me to identify my flaws and strengths, and for the patience to allow me to overcome the difficulties that appeared in my path during these years. I grew as an engineer and as a person. Thank you Jens for co-supervising me and thank you Paloma for the unceasing availability and willingness to discuss and help. A special thank you to Boeun, Brian, Ibrahim, Jan, Jay, Marta, Teresa and Thomas for the wonderful collaborations. I wish also to acknowledge my student Lauren Pickens who contributed to this project with laboratory assistance.

I am also thankful to my colleagues in PROSYS for their camaraderie. A particular thank you to Asbjørn, Christian, Franz, Leander, Mafalda, Morten, Rolf and Teresa for all the discussions, encouragement and smørrebrød. A Emílio, Johnny y Janet, gracias por la recepción cariñosa en México.

Lastly, but most importantly, a word of gratitude to my family, friends and above all to Sara for always being there for me. Obrigado.

Kongens Lyngby, February 28th, 2019

Ricardo Fernandes Caroço

Contents

CHAPTER 1

Introduction

The goal of this chapter is to contextualize the reader, to state the motivation and objectives which drove the development of this work and, give a brief overview of the case study employed. The primary focus of the work is set on tackling the effect of the natural variability of bio-based raw materials in processes, within an optimization and monitoring context, through the combination of model and data based soft sensors. Section 1.1 is dedicated to contextualizing the industrial setting of the contribution's impact field. Section 1.2 introduces the basic definitions and principles in which the contribution relies upon in an introductory fashion. Section 1.3 sets the discourse in the context of raw material influence motivating the work, as well as introducing a case study which is going to be used to illustrate the research objective that is stated in Section 1.4. An outline of the contents of the rest of the thesis is given in Section 1.5. Section 1.6 lists the main dissemination of research outputs.

1.1 Bio-based economy: a shift towards renewable raw material

The chemical industry has at hand a paradigm shift from petrochemistry to bio-based production. This is a change towards what is often referred to as bioeconomy or bio-based economy. The two terms, often used interchangeably, have nuanced differences in their definition, with bioeconomy referring to the biotechnological value-added products in the existing economy, and bio-based economy strictly related to

the production from renewable resources and their transformation into food, feed, bio-based chemical products, and bioenergy. These include the industries of agriculture, forestry, fisheries (aquaculture) and food production as well as sectors of the chemical, biotechnological and energy industries (Staffas et al. 2013). This is intimately related to the biorefinery concept, which is the process that uses biomass as a renewable feedstock to partially substitute fossil fuels for both productions of energy, fuels, and chemicals.

In 2012, the initial deployment of the bioeconomy strategy, given by the European Commission, had the primary goal of promoting a move from the current fossil-based economy (European Commission 2012). In 2015, the sectors encompassed by the bioeconomy accounted for circa 18 million jobs, which is approximately 8.2% of the EU's workforce. Between 2009-2015 the change in jobs in the bio-based manufacture of chemicals has been in the 10^4 order (European Commission 2018). Bio-based chemicals are expected to constitute the largest segment of potential growth for industrial bio-based products. Such bio-based chemicals can be produced either by bioprocesses or by conventional chemical processes using a bio-based feedstock.

Bio-based products also have a significant role in the transition from a classic to a circular economy. Furthermore, the waste generated on producing these bio-based products can be subjected to recovery, driving the bio-based economy further towards the 'zero-waste' society (Udugama et al. 2017). Significant focus is given to raw material, as it is the base for this transition. The availability, quality, robustness, and sustainability of a certain raw material affects its price, which in turn affects the economy of the process. The wastes from forestry and agricultural industries contain high value substances such as sugars, minerals and protein. The careless disposal of these wastes to the soil or landfill causes serious environmental problems, on top of constituting a loss of these added value substances.

1.1.1 Natural products

Natural products are part of these added value products. The group of molecules which can be described as natural products is very diverse in structure and can be isolated from many natural sources, making up to more than 200 000 already known and extracted compounds (Pearsall et al. 2015). Natural products can be alkaloids, coumarins, flavonoids, glycosides, iridoids, lignans, steroids and terpenoids, which are extracted and isolated from natural origins be it plants, animals, or microorganisms (Sarker and Nahar 2012).

These molecules are mainly divided into primary and secondary metabolites: the former are carbohydrates, amino acids, proteins, lipids and all other molecules which are synthesized for growth and development such as regulators, cell wall components and compounds composing the photosynthetic organelles like chloroplasts. Secondary metabolites are produced in synthetic side-routes besides the primary metabolites and are not part of the core molecules for a plant's survival. Their production can be associated with ecological responses throughout their evolution, for example in the

presence of predators or due to interspecies competition. Secondary metabolites are much more complex and diverse than the primary metabolite, from which they can derive from (Wu and Chappell 2008).

These secondary metabolites are of special interest if they have pharmacological (or toxicological) effects in humans and animals. If such is the case, they are called bioactive compounds (Azmir et al. 2013). These molecules have had a major role in the development of the food ingredient and medicine (pharma or natural) sectors. The development of new antibiotics, hormones and antitumor agents relying heavily (more than 60% from 1981-2014) on direct derivation from natural products, or have taken inspiration from it with slight semisynthetic modification (Newman and Cragg 2016).

Natural products from bioprocesses

There are different definitions of what the term *bioprocesses* comprises. A possible classification is suggested by Woodley et al. 2013, where the combination of substrate and catalyst is what defines the process. The processes considered to be bioprocesses can be assessed in Figure 1.1.

This approach considers that processing a renewable feedstock with conventional chemical catalysts, is a bioprocess. This encompasses the conventional extraction methods of natural products from natural sources (if renewable). The rest of the possibilities are biosynthetic approaches which use biological catalysts either by enzymatic bio-catalysis, fermentation or a combination of both.

For the commercialization of a natural product, the expansion to large scale production is necessary. Thus there is a need to develop these processes such that they have high quality and yield. Chemical synthesis of these molecules is difficult due to

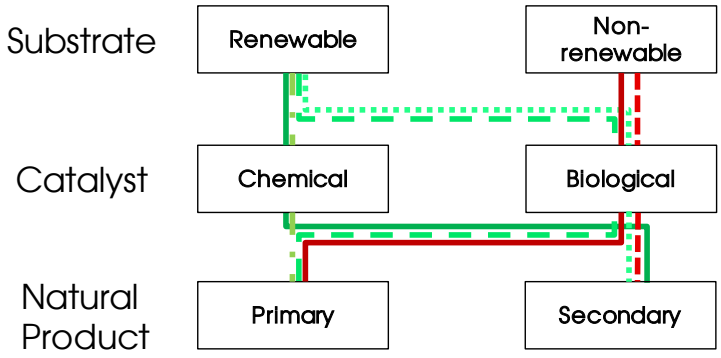

Figure 1.1: Different combinations of substrate-catalyst-product that are considered bioprocesses according to Woodley et al. 2013, in the context of natural products. In green color, are the combinations which are considered bio-based processes.

their complex structure and beyond the point in the biorefinery context. The direct extraction of products from natural sources is sometimes considered not economically feasible (Wu and Chappell 2008). The reasons for this are due to their relatively low concentrations *in planta* (especially of secondary metabolites), the presence of mixtures of different compounds and the variability of raw material. The raw material can vary in quality and quantity with climate and geographical changes. Therefore, biosynthesis has become a route exploited for the production of natural products, in the attempt of avoiding these limitations. Through metabolic engineering, the expression and insertion of genes responsible for the biosynthesis of a natural product in industrial micro-organism strains of yeast, fungi and bacteria, allows the production of natural products (Ajikumar et al., 2010) (Martin et al., 2003) (Anthony et al., 2009). Further use of biocatalysts (enzymes) can enhance this production.

In addition to conventional solid-liquid extraction which makes use of chemical solvents, represented in Figure 1.2, green extraction techniques (microwave, ultrasound, instantaneous pressure drop, supercritical fluid extraction, pressing) have been attracting attention and have shown to be a less wasteful alternative to the conventional solvent-based extractions (Rombaut et al. 2014). The fermentation in the scheme alludes to the submerged mode where the cultivation of microorganisms is performed in a liquid medium, which contains the necessary soluble carbon sources and nutrients. This requires a prior step of pretreatment (fractionation of the plant-based lignocellulosic material) to produce a fermentable hydrolysate. In contrast, solid-state fermentations are based on the growth of microorganisms on moist solid

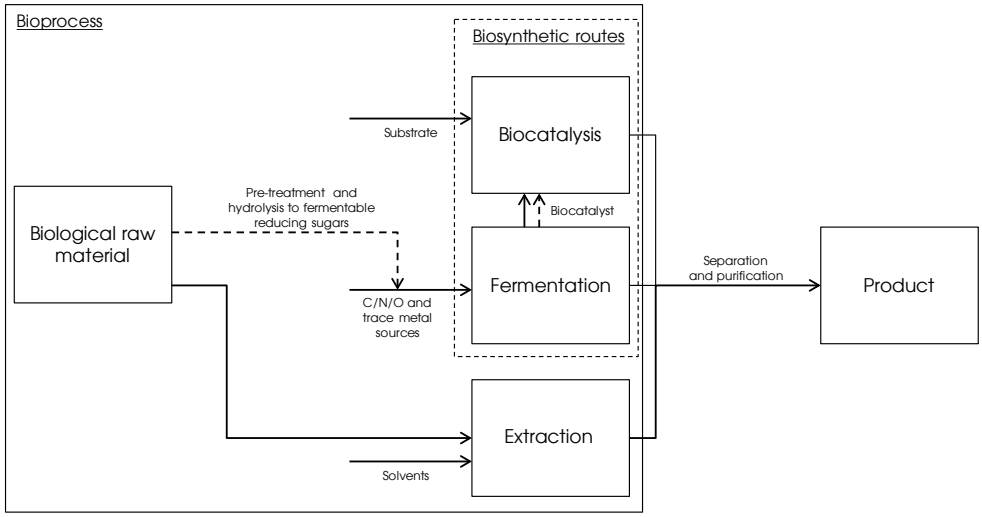

Figure 1.2: Diagram representation of the traditional routes for natural products manufacturing. The dashed arrows indicate the possibilities of using alternative bio-based feedstock in the biosynthetic routes

materials and have been gathering more attention due to higher yields and better product characteristics for the production of metabolites (Krishna 2008; Mussatto and Teixeira 2010).

1.2 M^{3}C - Measurement, Modelling, Monitoring & Control

Optimization, monitoring and control strategies are crucial in several stages of a product/process life-cycle, from the developmental phase to established commercial processes. These strategies are used to address a large number of objectives such as process understanding, troubleshooting, real-time control actions and continuous process optimization. M^{3}C is an umbrella term which encompasses the related fields which serve a role in these strategies. This denomination has been adopted by both the European Federation of Biotechnology and the European Society of Biochemical Engineering Science (C.-F. Mandenius and Titchener-Hooker 2013).

Regulatory initiatives

Significant efforts have been made by regulatory agencies in collaboration with different industrial and academic stakeholders in pushing the regulations and guidelines into a task of quantifying critical process variables directly in or at the process and (almost) in real time.

A key element to accomplish this task is process analytical technology (PAT). PAT can be understood as the combination of methods and tools which lead to process understanding and, ultimately, combined with measurements of raw material or in-process critical quality attributes allow for the control of the final product quality in real-time, by adjusting process parameters (Rathore et al. 2010). The US Food and Drug Administration formally proposed an incentive in the form of a guideline which calls for voluntary development and implementation of these and has been called the PAT initiative (Food and Drug Administration 2004). This position was cemented when the Q8 guideline was released and further advocated for the use of PAT systems, with the pledge of more flexible regulatory demands within this framework (ICH Expert Working Group 2009). This document also formally introduced Quality-by-Design (QbD), which is a supra-concept in which the PAT is one of the pillars.

QbD includes further concepts and tasks such as process knowledge management, process risk assessment/management, and operation design-space, which can overlap. These are concepts which have been developed before this report, throughout several industries. In contrast to traditional approaches, QbD emphasizes the characterization of the product's critical quality attributes. As a result, QbD-based control strategies become more pro-active and allow the process to be run flexibly, within the defined design space of a given product profile, to manage the incoming variability

in raw material characteristics. Thus reaching the desired product profiles is more likely and product quality can be better controlled.

QbD and PAT have been well accepted and adopted by the traditional pharmaceutical industry, where regulatory demands are greater. However, these processes are more straightforward (API chemical synthesis, isolation, stabilization, powder granulation and mixing, tableting, etc.). The awareness across the biotechnological and food industry has been growing in the past years, especially in the biopharmaceutical sector (Simon et al. 2015). The M^3C topics are a necessary foundation of a QbD-based approach for control, and they can be applied throughout the life-cycle of a process from early process development and understanding to the validation, monitoring, and control of commercial operations.

Measurement

The nomenclature used by Sonnleitner (2012) in the context of measurement systems is useful to establish a common ground. System *variables* refers to the properties of the system which are varying in time. Whilst, system *parameters* are those properties which are inherent to the system. The variables are thus the measurable properties of the system, whereas the parameters can be determined means of evaluation of the variables.

Measurement technology can be assessed from different classification perspectives as it can be seen in Figure 1.3. Regardless of type, the measuring system has to have a certain acceptable standard of accuracy, selectivity, and sensitivity with the desired key variables. The measurement system itself has to be robust, repeatable and stable. Furthermore, it is desired that the measurements have a low detection limit, are linear, have short response times and a have long use life (Sonnleitner 2012). Within the PAT-approach, spectroscopic measurements have a significant role as they enabled scientists to obtain rapid information about the process, and quantitative variable determination by means of chemometric modelling (Biechele et al. 2015).

Modelling

Models can be grossly divided in mechanistic and empirical. Empirical models represent input–output relations in a data set without requiring detailed knowledge of an underlying mechanism. Usually, an empirical model can only accurately predict conditions described by the data set that was used to build the model. Empirical models are useful in a process control context, where software sensors often rely on empirical models for the prediction of variables that are not measured directly owing to on-line measurement difficulty or excessive sensor cost.

Mechanistic models start from a conceptual understanding of chemical and physical processes (e.g., heat and mass transfer, chemical reaction rates, phase equilibrium, balance equations) that control a unit operation (e.g., reaction, mixing, separation). However, processes are usually represented by a combination of empirical and mech-

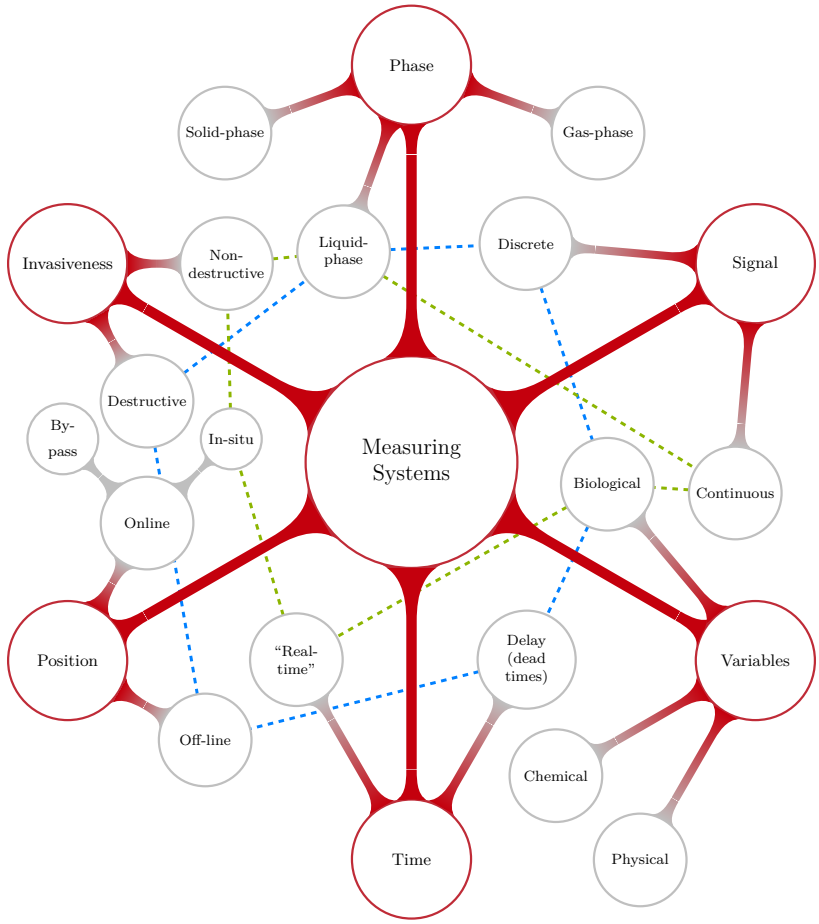

Figure 1.3: Different natures of measurement classifications. The dashed lines represent a classical HPLC laboratory analysis of the dissolved components in the media (blue) and a state of the art fluorescence probe (green)

anistic models, where mechanistic models gradually replace empirical models when more knowledge about a process or a unit operation becomes available (K. V. Gernaey et al. 2010). Specifically for the description of bio-based processes, the kinetic expressions themselves are often empirical, providing a simplified and idealized view of a complex biological mechanism.

Process models are useful for many different applications such as:

- Process equipment design
- Operation conditions optimization

- Variable estimation
- Parameter estimation
- Assisting process control

In the PAT context, mechanistic-based models offer greater process understanding as we can attribute physical-chemical significance to the estimated parameter, and also enable the analysis in terms of which parameters should be considered critical process parameters. Additionally, due to their better extrapolation properties it is possible to analyse other extended design-space possibilities for the process and can serve a central role in the real-time control/operation structure of the process.

Monitoring

Monitoring can be succinctly defined as the gathering of process data and posterior evaluation of process variables (or other parameters) to verify if the process is operating at the desired state of control. This allows for identification of process deviations enabling the implementation of corrective action, promoting continuous improvement. Once the important variables are identified and captured, the data from these variables are represented in the form of various charts and visualizations that can help in the detection of trends and deviations occurring in the process (Rathore et al. 2013).

Direct measurements of key variables are often missing or impractical in bioprocesses (Dochain 2008). Therefore, there is a fundamental need to develop a mapping of variables and parameters relationships through the use of models. This makes it possible to follow unmeasured key performance indicators of the bioprocess in real-time. These key performance indicators can be actual variables of the system or other mathematical constructs which allow the following of the process. Soft sensors have been established to enable indirect measurement that can provide access to relevant variables using mathematical techniques. This is represented in Figure 1.4. Soft sensors combine a measurement system (sensor) and a *software*-based algorithm for real-time monitoring. The measurement system direct outputs can be, as

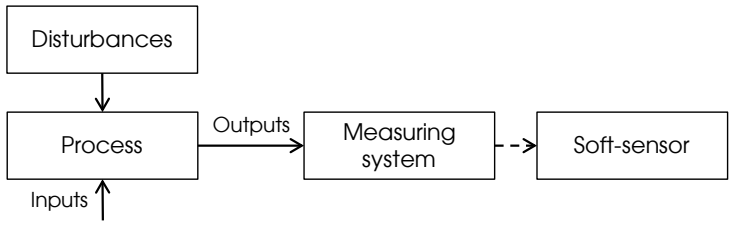

Figure 1.4: Schematic representation of a measured system with soft-sensor application

seen in Figure 1.3, of multiple formats. The algorithms allow for a transformation of the measurement outputs into the desired process variables that cannot be measured directly. This is referred to as variable estimation.

These sensors can be divided into two categories: model-driven and data-driven. Data-driven soft sensors are based on multivariate data analysis techniques, such as such as partial least squares (PLS), principal component analysis (PCA) and artificial neural networks (ANN) to describe the relationships between the outputs of the measuring system and the process variables. When using these data-driven approaches, one has to be aware of the calibration range. While the model-driven soft sensors are mainly based on the phenomenological models of the process and rely on how interconnections of variables are expressed in the system model, as well as how the measurement is related to these variables (Sagmeister et al. 2013a). The sensors are used as appropriate alternatives and/or additions to standard hardware sensors to meet PAT and quality by design requirements (Luttmann et al. 2012).

Control

Control structures can be feedback or feedforward in mode, irrespectively of the control law algorithms that we choose. Both are represented in Figure 1.5. In closed-loop control, the objective is to minimize deviations in the controlled variable from a defined set-point and the algorithm necessarily uses the system's output measurements, which might require the pre-processing by means of a soft-sensor, for the actuation effect on the related manipulated variable. Feedforward makes use of input disturbance measurements for correcting their effects on the key controlled variables. The main disadvantage of the latter is that if there are internal variations in the system, there is no in-process correction.

It is then advantageous to, whenever possible, have control laws which use mixed structures with measurements in the input disturbances and system output. This

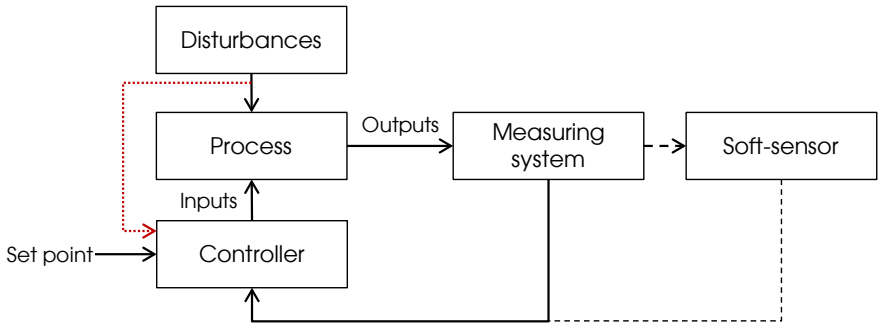

Figure 1.5: Schematic representation of a controlled system. The dotted red arrow from the disturbances to the controller represents the feed-forward control possibility.

allows for a pre-emptive action based on the input disturbances, which in turn minimizes the influence of these disturbances on the outputs making it possible for the feedback to better compensate the internal system variation (Dochain and P. Vanrolleghem 2010).

The effective automatic control strategies implementation is the last piece in a PAT-based approach that renders control loops revolving on the quality of the product, leading to an optimized real-time operation or recipe development for specified product critical quality attribute profiles (C.-F. Mandenius and Titchener-Hooker 2013).

1.3 Motivation

The main goal of the study is to improve processes which convert biological raw material by developing tools for measurement, optimization, and monitoring which fit in a model-based operational methodology.

Many industrial processes with this type of raw material rely on feedback from historical production performance to withdraw insights to improve the process for a given type of raw material. This could be wasteful in resources and time-consuming, with no guarantee of systematic optimization.

Raw material quality fluctuation represents an important element of disturbance in many industries. More specifically, those that use a diversified supply of biological raw materials. The production subjected to uncontrollable variation due to a multitude of reasons, such as differences in quality across different suppliers from different origins, and even the yearly and seasonal variations within the same supplier. Hence it is valuable to exploit effective ways to analyse raw materials and to adjust for the corresponding variation in the process, in order to succeed in the production of products within their specifications. As seen in Figure 1.6, increasing interest in the bio-based economy in the literature is accompanied with also a growing interest in the quality (and variability thereof) of raw material. In fact, an ultimate report from an *Expert Group on Bio-Based Products*, which reviews the previously mentioned 2012's Bioeconomy Strategy and Action Plan, states that

> *"The main driver of innovation in the bio-based sector is to gradually ensure that all potential feedstock sources are processed in a manner that uses all components in the smartest, most efficient way, whilst adding the greatest possible value"* (European Commission (2017))

There is then a need for an adequate operational approach that can mitigate production performance issues and undesired deviation of the critical quality attributes of the end-product due to raw material quality fluctuation. The development of successful dynamic models that combine process condition and raw material related parameters enables a tool for optimization and monitoring purposes for the incoming

feedstock. For a given desired product specification, having information on raw material which is incorporated into a process model allows for the optimization of the process conditions. This allows the use of the biological raw material, to the fulfilment of the targeted production efficiently. Additionally, it will enable the manufacturer to predict if certain target productions are possible for a given raw material.

This type of models, based on mechanistic and dynamic phenomena, are advantageous in fullscale application versus the purely empirical methods, but also requires a more solid understanding of the root causes of the phenomenon. These models are also practical for simulation studies at scales and condition different than the ones from model calibration. The mismatch between the model and plant behaviour can be reasoned and interpreted (i.e., different transport phenomena rates than at model development scale) and corrected throughout a batch for the online prediction and monitoring, by means of model-based sensors.

The combined approach of characterizing raw material and modelling the process key performance indicators in an intertwined solution, provides the manufacturer for more operational flexibility and is a step towards ensuring feedstock use efficiency. An approach which is able to predict the quality at the process end time also enables the production to the desired specification. This also changes the current typical practice of mixing and blending product at the end of the line to obtain the required specification and opens opportunities in terms of raw material and product inventory management.

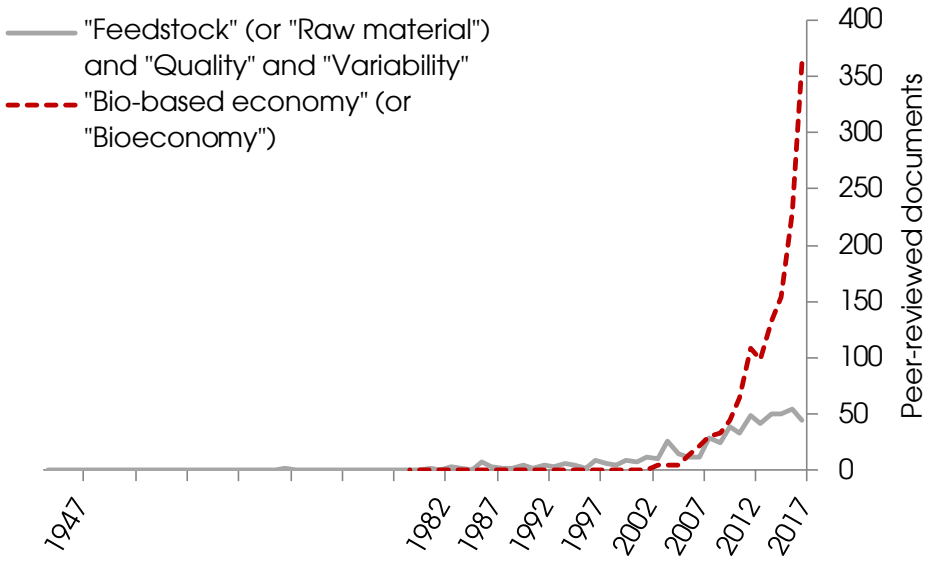

Figure 1.6: Number of publications per year by titles, abstract or keywords search terms in the Scopus document database

The pectin solid-liquid extraction case study provides an example of combined use of input assessment and process models for optimization and monitoring in the context of a bio-based (raw material) converting process.

1.3.1 Case study: Pectin extraction from citrus peels

This research project has been financed by Innovation Fund Denmark through the BIOPRO2 strategic research consortium which counts with the participation of major biotech key players in Zealand, Denmark. This project was conducted in close collaboration with CP Kelco, one of the stakeholders of this consortium. CP Kelco is a hydrocolloid producer, with a strong emphasis in the manufacture of pectin molecules for application in a diverse portfolio of consumer products.

Pectins are primary natural products, of anionic polysaccharide structure, extensively available at the cell walls of fruits. Structurally they can be divided into three domains: homogalacturonan (HGA), rhamnogalacturonan-I (RG-I) and rhamnogalacturonan-II. The commercialized hydrocolloids commonly are constituted by blocks of HGA and RG-I domains. HGA domains are linear sequences of partly methyl esterified α (1→4) linked D-galacturonic acid residues (forming a "smooth" backbone). Whereas in RG-I domains this backbone is interrupted by α (1→2) linked rhamnoses where side-chain groups attach, mostly composed of β (1→4) linked D-galactose or α (1→5) linked L-arabinose (Sriamornsak 2003; Willats et al. 2001).

These polymers have been predominantly used in the food and beverage industry due to their gelling and thickening properties. For different product applications, pectins with distinct critical quality attribute profiles are required. These critical quality attribute profiles are composed of the different physical and chemical characteristic combinations, which pectin can display. The degree of esterification ($\%DE$) and the intrinsic viscosity (IV) are critical quality attributes for the gelation properties of the pectin product.

Industrial pectin manufacturing consists of extraction from the plant material (citrus fruit peels), purification of the liquid extract, precipitation of pectin from the solution, and further de-esterification of high methyl ester groups with acid, alkali, and ammonia depending on the desired pectin quality profile. The extraction step is the one which deals with the biological material as inputs, and it runs as a maceration (batch process) in hot acidified aqueous solution for several hours. Pectin diffuses through the previously washed and dried citrus peel to the bulk solvent liquid. Insoluble pectin, called protopectin, is solubilized in the presence of a dilute mineral acid (HCl, HNO_3 or H_2SO_4). The desired critical quality attribute profile is reached by running at a specified pH (1.5-3) and temperature (50-100 °C), for a given extraction time (Ciriminna et al. 2016).

The market has experienced an increase over the past few years, being reported that the global pectin market reached $1 billion in 2015. This trend was expected to continue over the next years due to the increasing application scope of pectin in diverse industries (Ciriminna et al. 2016). Pectin is seen as an emerging bioactive compound

(Maxwell et al. 2012). More applications in different sectors (i.e., pharmaceutical, medical and cosmetic) are being investigated, which would further increase the market value (Lebreton-Decoster et al. 2011; Lupi et al. 2015; Sriamornsak 2003).

The scope of this project is interesting for CP Kelco since the increase in pectin demand contrasts with the scarcity, variability and rising prices of its biological raw material. This bottleneck in supply can lead to customers, in all markets, search for alternatives due to necessary price hikes to follow-up the supplier's inflation. It has been reported that pectin price increases in a range 10-30%, in consecutive years, was directly attributed to this cause (Gray 2014; Scott-Thomas 2013). With this challenge, the industry needs to face a paradigm shift from a heuristic recipe-driven operation to a more informed, lean, and flexible operation regarding the incoming raw material since, unlike other process conditions, the manufacturer does not directly control raw materials.

1.4 Research objective and project goals

The thesis main research objective is to provide:

- A systematic model-based operational strategy for processes with biological raw material as a disturbance.

The operational strategy enables an optimized recipe for the production of the desired quality target, given a characterized input raw material. The strategy is completed with the incorporation of in-process measurements that are used to update the model, to recast prediction to assess if the plant is reaching the defined product target and from there, take corrective actions.

This main goal is composed of individualized research topics, of which their individual goals in this project are:

- To develop a systematic raw material characterization approach, with the purpose of utilizing the information to link with the process model for process optimization and monitoring. An approach which can rely instead on faster and less wasteful spectroscopic tools combined with chemometric techniques, to prevent having to rely on the production feedback approach to assess a practical definition of quality of raw material.

- To analyse existing extraction models in terms of the effect of raw material and other model parameters.

- To develop tailored process optimization for different end products for a given raw material while using a process model that combines raw material and process conditions variables.

- To exploit the model to construct soft-sensors and state estimators for monitoring and correcting predictions of variables of interest throughout the production

series, allowing if need, a better informed process conditions update for the following batches.

In the context of the pectin case study, this culminates in improving the pectin manufacturing recipe in terms of temperature, pH and time, to reach the desired degree of esterification ($\%DE$) and the intrinsic viscosity (IV) profiles, while maximizing extracted pectin concentration. These variables are then assessed throughout the extraction time via the use of model-based algorithms and discrete sample data. These four objectives are connected and revolve around the pectin solid-liquid extraction process as represented in Figure 1.7. All published data and available models for pectin extraction, by the time of this project, dealt only with laboratory and pilot scale measurements.

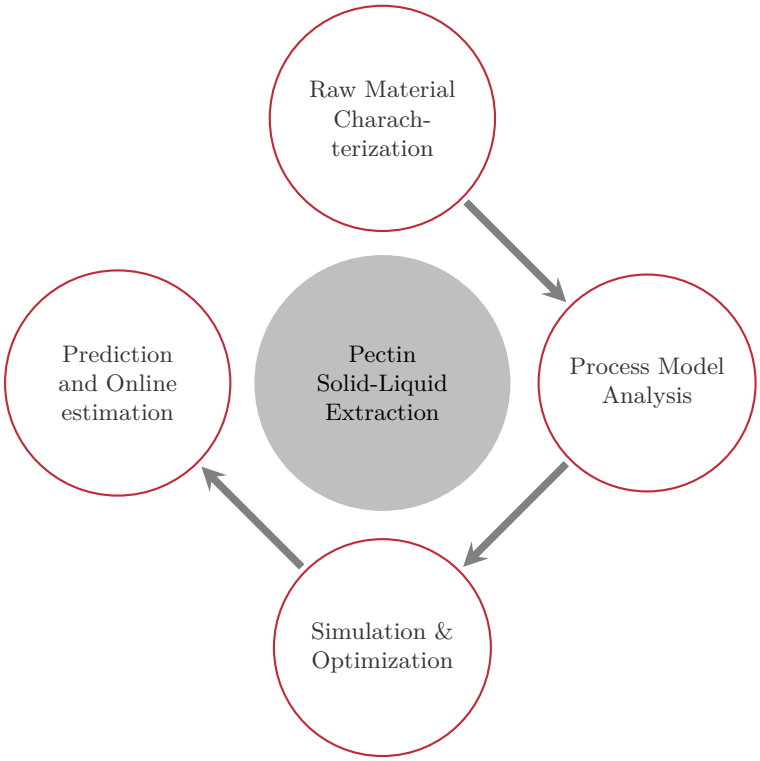

Figure 1.7: Diagram of the addressed topics in relation to the case study.

1.5 Thesis organization

The thesis is divided into three parts composed of chapters, excluding this introductory chapter, conclusion, future perspectives, appendices and references, as follows:

Part I contains the state-of-art on model-based aspects of extraction, which then leads to the definition of the proposed approach for an optimized operation scheme.

- **Chapter 2** starts with a review of existing practices the use of different models and model-based process optimization and monitoring approaches for extraction processes. It also provides a particular review of the pectin case study.
- **Chapter 3** presents the systematic workflow to develop an operational strategy for processes which are affected by biological raw material disturbances.

Part II showcases the individual results of each individual objective of the main contribution, in the context of the introduced case-study.

- **Chapter 4** presents the development of the raw material characterization approach for citrus peels. The modelling aspects of developing a PAT tool based on near infra-red spectroscopy are discussed and compared with different strategies for raw material quality assessment.
- **Chapter 5** deals with the modelling and respective model analysis of the process. A mechanistic approach is investigated, based on the prior work of Andersen et al. 2017. This pilot-scale model is investigated for parameter sensitivity and additional uncertainty analysis is conducted in order to assess the model limitations.
- **Chapter 6** provides the strategy for recipe optimization based on the developed model. The optimization provides set-points for temperature, pH and extraction time depending on the desired product profile and incoming raw material.
- **Chapter 7** presents the development of a tool to forecast the process key performance indicator variables based on the process model coupled with the prior raw material information and in-process measurements, relying on state and parameter estimation algorithms.

Part III presents the results of the full operational approach and provides discussion on its application.

- **Chapter 8** delivers the combined implementation of the previous sections, ultimately showcasing the proposed operational strategy in the context of pectin extraction for a particular citrus peel.

The thesis is finalized with a conclusion chapter and with a list of prospective ideas and future development from this work.

1.6 Dissemination

All the scientific dissemination produced as part of the research conducted in this Ph.D. study are listed and their relation to the thesis chapters is highlighted.

Journal papers

- Caroço, R.F., Bevilacqua, M., Armagan, I., Santacoloma, P., Abildskov, J., Skov, T., Huusom, J.K. (2018). "Raw material quality assessment approaches comparison in pectin production". In: *Biotechnology Progress* [**Chapter 4**]
- Caroço, R.F., Kim, B., Santacoloma, P., Abildskov, J., Lee, J.H., Huusom, J.K. (2019). "Analysis and Model-based Optimization of a Pectin Extraction Process". In: *Journal of Food Engineering* 244, pp. 159-169 [**Chapter 5 and Chapter 6**]

Peer-reviewed conference submissions

- Advances in Process Analytics and Control Technologies - APACT 2016

  "Model Based Monitoring of Bioprocessing Plants: A Solid Liquid Extraction Example".

- 18th International Conference in Near-Infrared spectroscopy - ICNIRS 2017

  "Spectroscopic Raw Material Fingerprinting As An Early Stage Step In A Monitoring Strategy For Industrial Pectin Extraction"

- Nordic Process Control Workshop 2016

  "Model-Based Monitoring of an Industrial Batch Pectin Extraction".

- Nordic Process Control Workshop 2018

  "Comparison of two classes of observers in a biochemical process"

Part I
Improving Industrial Product Extraction from Natural Sources

CHAPTER 2

Conventional Industrial Extraction

This chapter provides a state of the art foundation and it serves as the background knowledge for what led to the proposed operational strategy in the succeeding chapter. Section 2.1 describes the current existent practices in the conventional industrial extraction of valuable substances from natural raw materials. Section 2.2 presents the state of the art in terms of modelling practices in this field, referencing the work done by different researchers. Section 2.3 examines the existing methodologies in design, optimization and monitoring that incorporate extraction models systematically. Section 2.4 is dedicated to the pectin extraction process.

2.1 Extraction of products from bio-based materials

The direct extraction of products from biomass is a millenarian activity, having a core role in traditional medicine of different ancient civilizations (Newman et al. 2000). This practice has been throughout the years extensively experimented and studied from the standpoint of discovering new natural product molecules from different raw materials.

The typical raw material are plant and other lignocellulosic materials, which are available from the forestry, milling, and agricultural industry residues. They are mostly made of cellulose, hemicelluloses, and lignin. From these renewable natural materials, there is a wide variety of products that can be obtained through many different methods. The different process alternatives for recovery of these products are intimately dependent on the properties of the targeted active component. These methods are typically within the scope of cold-pressing, distillation, steam and hydro-distillation, and solid-liquid extraction methods (Bart and Pilz 2011). This chapter focuses on the latter method.

Some classic examples of long-standing successful industrial implementations of solid-liquid extractions are the colour pigments and dyes (Jothi 2008; Sivakumar et al. 2011), sugar (sucrose) from sugar beet and sugar canes (Both et al. 2013; Brüniche-Olsen 1962), lipid and oil products from seeds (Langhurst 1951; Le Clef and Kemper 2015), caffeine from coffee beans (Bichsel et al. 1976; Chiang et al. 2018), or functional hydro-colloids from algae or fruits (Ciriminna et al. 2016; May 1990; McHugh 2003; Rhein-Knudsen et al. 2015).

2.1.1 Solid-liquid extraction process

Solid-liquid extraction (also commonly named as leaching) is a process that can be applied across several industries in the chemical and biochemical processing realm.

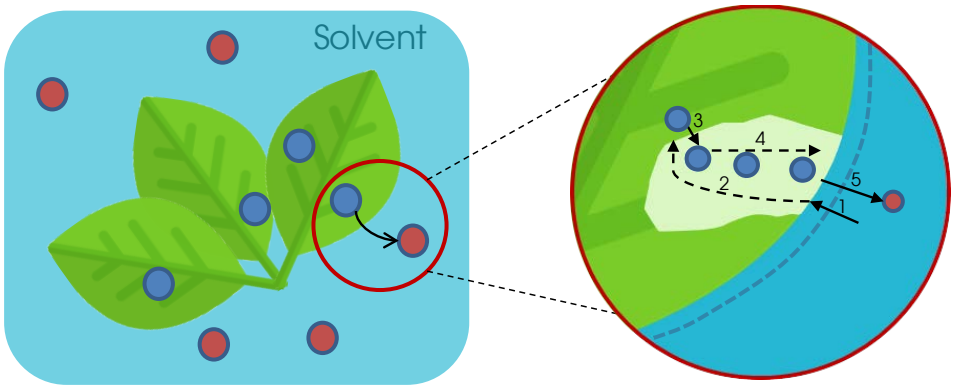

Figure 2.1: Representation of the steps involved in extraction

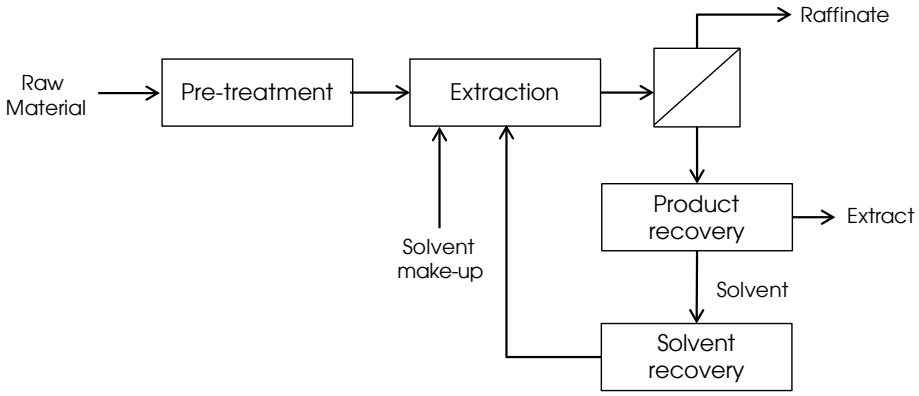

Figure 2.2: Schematic representation of the integrated extraction process

In its essence, it can be simply described as the removal of a soluble fraction from an insoluble solid phase, by contact with a solvent phase. A pictorial representation can be seen in Figure 2.1.

The solute concentration difference between the liquid and solid media is the driving force for the diffusion from the raw material to the solvent (Abidin et al. 2013). However, it can be observed that the process is reliant on other different stages. In Figure 2.1, step 1 and 2 are related to the penetration of the extracting solvent in the pores of the raw material (i.e., solvent convective mass transfer from bulk to the surface and diffusion through the pores of the raw material). Step 3 depicts the dissolution or solubilization of the desired solute, which may involve a reaction of a precursor molecule. The rate of the extraction can thus be dependent of underlying reactions, which affect the overall kinetics of the process. The solute is diffused through the pores of the solid material (step 4) and then transfered to the bulk liquid which envelops the raw material (step 5). The process continues until an equilibrium of concentrations in both phases is reached (Seader and Henley 2006).

Prior to extraction, raw-material pretreatment is often necessary to allow an easier penetration by the solvent. A classic approach is to grind to smaller particles, which provides an increased contact area for solvent impregnation and, in many applications, breaks cell structures. Other methods like freezing, thawing, swelling (with water or organic solvents) are proven to provide higher yields in the succeeding extraction, however the application depends on the application, as it is essential to not cause degradation of the desired compounds. The extraction occurs when the solvent is mixed with the solid material. The outcome of the solid-liquid extraction is separated into a stream containing the residue solid material (raffinate) and a stream which can be made up of a complex multicomponent mixture that require subsequent purification steps, to achieve the desired pure substance. Solvent recovery is an important part of the integrated process, since great amounts of solvent are necessary

and its recovery is needed for the economic feasibility of the production. This steps are represented in a simple block diagram in Figure 2.2.

2.1.2 Extraction technology

The solid-liquid extraction process relies greatly on the solubility of the target solute in the solvent at the proposed process conditions and raw material structure. Hence, the kinetics are mainly influenced by parameters such as solvent choice, pH, temperature, solid–liquid ratio and particle size. Extraction technology withstood an evolution, which led to many different equipments and apparatus. The different extraction variations can be identified depending on their features. Figure 2.3 summarizes the possibilities within the different categories, of which a given solid-liquid process belongs to. The choice of the type of extraction system depends on the raw material as well as the quality target profile of the desired compounds.

Principles

The type of contact between the solvent and solid matrix can be distinguished in percolation or dispersed solids, which is further differentiated in immersion and maceration techniques. Percolation extractions involve fixing the solid matrix as a fixed bed, while the solvent passes through. The solvent may be circulated multiple time

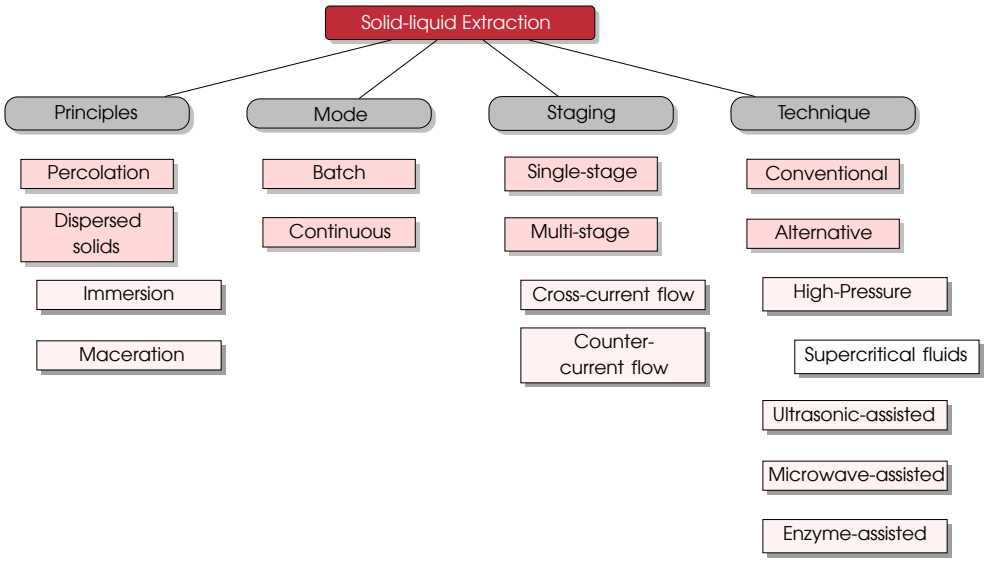

Figure 2.3: Diagram of the different solid-liquid extraction categories and their different possibilities in the context of extraction from bio-based materials

through this bed to increase contact time until the required extraction mass has been achieved. In dispersed solids extraction, the solid and solvent are always in full contact. The raw material is submerged in the solvent. In immersion, the contact is ensured by agitation/stirring, whilst maceration simply implies soaking the solid materials with no input motion. Other nomenclature variations exist, but they are attributed to processes which fit in both categories, yet having some recognized specificities such as, for example, the immersion followed by a cooling of the mixture being referred to as an infusion or the immersion in a boiling liquid solvent being known as decoction.

At laboratory scale, the standard analytical set-up is the Soxhlet apparatus which operates on a percolation principle. The apparatus consists of an solvent collection flask coupled with a sample chamber and a reflux condenser. The raw material is fixed in the middle sample chamber above the solvent in the flask. The solvent is heated to reflux and then then passes a condenser at the top of the extractor flask and as is it condenses the solvent is rinsed back into the solvent collection flask. As a result, the plant material is extracted several times with fresh solvent and the extracted compounds are collected in the bottom flask. Throughout the Soxhlet method the evaporation, condensation, extraction and reflux of solvent are repeated until the concentrated extract can be obtained from the collecting flask. However this method is not adequate for all compounds, as the solutes are in contact with solvent at its boiling temperature, which may be unwanted for thermo-labile compounds (Bart and Pilz 2011).

The choice between both principles may reside on the physical constraints of the raw material, which might not be suitable to withstand the shear velocities of solvent flows in percolation. However, percolation is more suitable for solids with low internal resistance to mass transfer, and immersion should be favoured when the resistance to diffusion inside the particles is significant (Abidin et al. 2013). Percolation has the added benefit of requiring less downstream separation units (e.g., filtration, centrifugation).

Mode and Staging

The mode of operation is a major classification of the type of extractor used in industrial scale. It is also intimately related with the staging category. Batch mode is the simplest and was the norm for traditional extraction products. In batch mode, solvent must be in contact with the solid until equilibrium is reached, while in continuous mode solvent and solid material are continuously inputs while recovering the extract and removing the raffinate (Meireles 2009).

In a single-stage batch percolation, fresh solvent continuously flows through the solid material which is getting depleted of the desired solutes. This use of fresh solvent promotes a higher driving force for extraction, but is also a more wasteful compared to a single immersion batch. This has led to the operation of multi-stage batch operation, using a series of batch tanks. Connecting the effluent of a percolation extractor to a second extractor, once its effluent drops below a certain concentration threshold, and

consecutively to the following tanks is a commonly used scenario. Fresh solvent is only fed to another tank, once the solid of the preceded tanks is depleted. This can also be referred to as pseudo-continuous or semi-batch modes, since the solids are immobilized in each stage but the solvent feeds are flowing between stages (Abidin et al. 2013). Another multi-stage batch scenario is by performing washing steps posteriorly of the main extraction step and re-utilize the effluent for the incoming batch extractions. These approaches improve the extraction yield and allow for better solvent utilization. The multi-stage batch process can be modified to run at optimized extraction times and to fit required capacity in production (Bart and Pilz 2011).

On the other hand, continuous processes have been increasingly implemented due to their improved efficiency in terms of extraction yields and solvent and energy consumption. Counter-current flow, as in many other processes, is the most efficient continuous mode which essentially requires that a fresh solvent stream is the input of the extractor containing the most depleted solids, and the extract of this stage is passed through a series of extractors successively until it reaches the extractor with the fresh raw material loading (Abidin et al. 2013). This promotes a high concentration difference between the solvents and the solid in operation. The cross-current mode implies the use of fresh solvent at each extraction stage, and albeit promoting the total extraction of the solutes it also demands the biggest amount of solvent.

In industrial practice the equipments and production schemes are often combined to serve the specific purposes of the extraction in terms of yield, product quality and capacity, for a given raw material and desired product profile.

Technique

The extraction is traditionally performed making use only of an organic solvent, heat and agitation. The solvents used are dependent on the compatibility with the nature of the desired bioproduct, as well as, economic, safety and regulatory aspects. The crucial points of concern in the choice of solvent are the solubility of the compound (polarity) to extract, the selectivity for its extraction as well as the stability and boiling point of the solvent. Other physical properties which also have a role in the solvent selection due to their importance in the integrated process scheme (Figure 2.2) are viscosity, inter-facial tension and the vaporization and latent heats. Typical solvents include ethanol, hexane, acetone and water (Bart and Pilz 2011). Alternative techniques have been developed with the aim of reducing extraction time, solvent consumption and better selectivity.

Figure 2.3 also includes the most represented forms of alternative extraction techniques. High pressure extractions with supercritical fluids (commonly called supercritical fluid extractions) are the most prevalent extraction alternative applied in the industry. In this supercritical state, the solvents have extremely low densities, which lead to an impregnation into the solid material with ease, and the high pressures lead to higher solute solubility. Supercritical CO_2 is often used as solvent due to its low critical point (31 °C , 7.3 MPa), availability, and safety (non-flammable and non-toxic). Other type of pressurized extractions are common solvent extractions performed at

elevated temperatures (100-200 °C) and at pressures between 10 and 15 MPa, with the solvent in liquid phase (commonly called accelerated solvent extraction). The high temperatures and pressure promotes a faster extraction by forcing the liquid into the solid matrix and enhance diffusivity of the solvent. Ultrasound technology has also been tested for the enhancement of extraction rates and yield, by using waves in the 18–20 kHz range to disrupt the solid material structure. Microwave-assisted extractions have the goal of reducing solvent use and extraction time through radiation, typically in a 0.896 - 2.45 GHz range. The radiation provides a more homogeneous heating of the solvent and solid matrix. This could also lead to rupture of the internal structure of the solid material, allowing a facilitated mass transfer. Extractions which make use of enzymes as catalysts have also been gathering attention. Enzymes have the capability of catalysing reaction in very selective ways, and perform under mild aqueous conditions. They can degrade cell walls and membranes, enabling more efficient extractions of the desired compounds.

All these options follow the research tendency to provide *green extraction* operations (Azmir et al. 2013; Bergeron et al. 2012; Chemat and Cravotto 2013; Joana Gil-Chávez et al. 2013; Marathe et al. 2017; Puri et al. 2012; Reverchon and De Marco 2006; Rombaut et al. 2014; L. Wang and Weller 2006). However these advances have mostly been applied at laboratory scale, as the constraints in equipment design have hindered its scale-up and application industrially. The most popular methods at an industrial scale remain to be supercritical fluid extractions and conventional maceration and immersion extractions (Abidin et al. 2013; Kassing et al. 2010b; L. Wang and Weller 2006).

2.2 Process modelling

Modelling is a systematic way to reproduce a system's response to a given set of input and disturbances. Such representation of a solid-liquid extraction system can be made by describing the phenomenological principles which govern the process (i.e., mass transfer phenomena and equilibrium conditions), empirical equations which represent the extraction kinetic dynamics or rely on the development of statistical models that are based on experimental design (DoE).

The models attempt to describe the solute quantities which are extracted to the surrounding solvent over time. A typically observed extraction curve for a batch system represented in Figure 2.4. It comprises a extraction stage, followed by and equilibrium stage. The extraction stage can further be divided into a fast extraction step, commonly named as *washing*, and a slower extraction step, commonly referred to as the *diffusion* stage. It starts with the external process of washing at the surface level (external process of washing), where the solutes are dissolved in the solvent and are immediately transported to the surrounding solvent medium. Following this, the successive steps represented in Figure 2.1 come into play, in which the slowest process is typically the diffusion step, hence the stage name. This concentration versus time

data are very commonly measured with the purpose of being fitted to models that, under appropriate hypotheses, allows the determination of the kinetic parameters.

Influential Process Parameters

The main process parameters which affect the conventional solvent extraction systems are the extraction solvent, solvent to feed ratio, raw material particle size and extraction temperature (Chan et al. 2014).

The choice of type of solvent (and its concentration) is an important process parameter of an extraction. It can influence both the speed of the extraction and its extent. Not only the diffusivity of the solute in the solvent is important, but also the solvents ability to penetrate the solid matrix.

The ratio of the solvent to feed is also a tunable process parameter which at its optimum decreases the mass transfer barrier in the diffusion stage. Over this optimum, the operation is wasting solvent. The washing stage is not usually altered by changes in this process parameter, but lower ratios will lead to lower yields at the end of the extraction.

Particle size is a process parameter which greatly affects the kinetics of an extraction. As with the solvent to feed ration, the washing rates are usually not affected, but the diffusion step rates are improved (Herodež et al. 2003). This improvement in mass transfer is due to the larger solvent-solute surface contact area and shorter diffusion paths (Hojnik et al. 2008). However, this effects are dependent on the solid

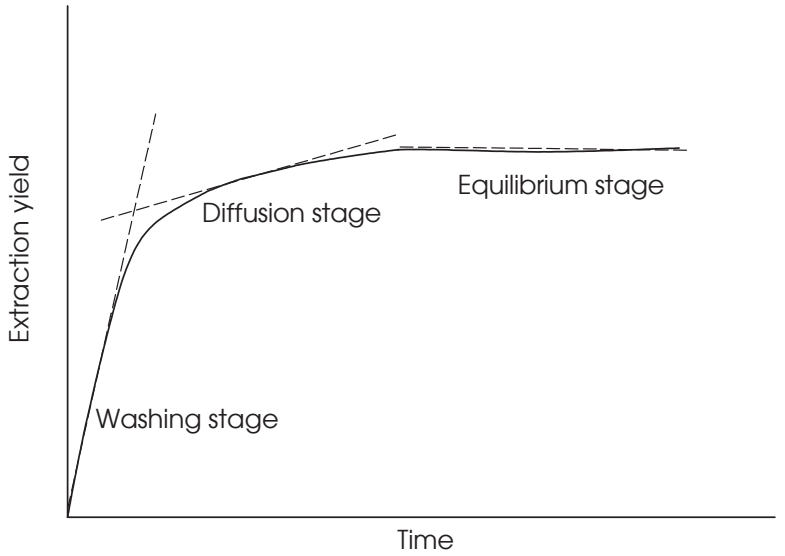

Figure 2.4: Extraction profile curve representation for batch extraction systems

matrix geometry and it has been seen that for plate-like samples the effect of particle size was not relevant, since the mass transport dimension is thickness (Wongkittipong et al. 2004).

Temperature affects the extraction of compounds in multiple ways. It decreases the viscosity and surface tension of the solvent, which increases its solvating power. The solvent has better capabilities of solubilizing solutes, and the matrix wetting and penetration is also improved (Chemat and Cravotto 2013). Faster diffusion and higher yields are thus obtained at higher temperatures. However, for thermo-sensible compounds this increase has to be regulated depending on the stability of compounds properties at high temperatures.

2.2.1 Empirical modelling

Empirical modelling is an important tool for process optimization, which allows for the optimization of processes without having the need of a fundamental knowledge of the process intricacies. However it has the caveat of being specific for the given system and not transferable to further systems. This leads to cost and time intensive experimental efforts if the design space to be investigates is vast.

2.2.1.1 Statistical modelling via Design of Experiments

Design of experiments (DoE) is a powerful method for the planning and interpretation of data that is extensively used in engineering problem solving, with the ultimate goal of increasing process knowledge and ultimately optimize product quality and process efficiency. It is an extension to the classical one-variable-at-a-time (OVAT) analysis strategy, in which the experimenter varies one variable while keeping other factors fixed. This univariate approach may produce false optimum results for a process, as it does not account for possible joint effects of different variables on the process response (Box and Draper 1987).

A DoE approach makes a tentative approximation of a real complex problem by means of a generic mathematical model in order to understand the relationship between a set of inputs and outputs of a process. In DoE terminology, factors are the operational variables that can be manipulated (X) that are either quantitative/continuous (temperature, pH, solvent concentrations, pressure etc.) or qualitative/discrete (type of raw material, type of supplier, type of additive, etc.). Additionally, variables that are difficult to manipulate are named uncontrollable factors (operators, room temperature, etc.) and an effort to mitigate them is crucial to minimize the process inconsistency. The dependent variables (Y) are referred to as responses, and deemed as continuous (Antony 2014).

The starting point for the selection of the type of design is the definition of a experimental objective, selection of factors and range for their variation, in which a "level" is referred to a specific setting of the factor being tested and a "run" is a the combination of factor levels to whose effect on the response we want to assess. The key

factors are identified either from past experience or through screening experiments. Reducing the amount of factors to be assessed is important to reduce the amount of experimental runs. The experiment expenditure requires studying at least three levels (per factor), for estimation of the regression coefficients, which for a large number of factors increases runs exponentially.

A full factorial design that contains every possible run (3^N) to be employed for 4 factors, requires a total of $3^4 = 81$ runs. Full-factorial designs may lead to a prohibitive number of runs, hence fractional designs and other alternatives are typically employed, providing less time (and resources) consuming designs. Re-do runs and additional runs to strengthen the model lack-fit have also to be taken into consideration.

An experiment is planned accordingly to its objective and once the key factors of a process are identified, response surface methodology can be performed if the objectives are to obtain the optimal response (and process settings), look for the *weak* response areas and identify the settings of the process which are less prone to perturbations. Each experimental response (Y), is described by a second order polynomial model that predicts the response in the experimental design regions through the following equation:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \beta_{ii} X_i^2 + \sum_{i=1}^{n-1} \sum_{j=1}^{n} \beta_{ij} X_i X_j \tag{2.1}$$

where Y is the desired modelled response, by means of interaction of the X factors (parameters). The linear, quadratic and cross product coefficients β_i, $\beta_i i$ and $\beta_i j$, as well as the constant β_0, are estimated via linear least-square regression.

This model provides a response surface and can be used to predict the responses for the complete experimental range of the factors. Furthermore, the estimation for all possible factors combinations allows us to find the optimal response and, consequently, where the optimized process settings are situated. The response is usually represented graphically in three-dimensions or through a X_i and X_j plane with drawn curves of constant response, called contour plot.

This type of statistical modelling is a common method for optimizations and has its advantages in identifying the interaction between different factors and their influence on target outputs, without necessary knowing the fundamental cause (Antony 2014). Experiment-based studies for defining chemical and biochemical processes optimal operational settings are a common practice (Bas and Boyaci 2007).

2.2.1.2 Kinetic empirical models

A different empirical approach for modelling the extraction behaviour relies on employing different mathematical equations which match the extraction curve. The adjustable parameters of these equations have thus no physical meaning and is not possible to extract any mechanistic understanding of the process from their estimation, but can be related to some extent to the known process phenomena. Below

are some examples of these models that have been employed in the context of batch extractions.

Peleg

The model developed by Peleg (1988) for the description of sorption curves has been adapted to the solid liquid extraction dynamics as follows:

$$c(t) = c_0 + \frac{t}{K_1 + K_2 t} \tag{2.2}$$

where K_1 is the model rate constant, K_2 is the model capacity constant and c_0 is usually equal to zero. The Peleg rate constant (K_1) can be related to the kinetics at the very beginning of the extraction ($t = t_0$) while K_2 relates to maximum of extraction yield, i.e. equilibrium concentration (Jokic et al. 2010).

Rate Law

The rate law can be adapted to model solvent extractions, by use of the second-order rate function (Chan et al. 2014). The equation is given as follows:

$$\frac{dc}{dt} = k(c - c_\infty)^2 \tag{2.3}$$

where k is the second order extraction rate constant. This approaches assumes that the solvent concentration tends to a constant value of c_∞ for a given set of conditions. The solvent is assumed well mixed, with uniform concentration throughout the extraction.

The parameters can be estimated by linear regression to experimental data of the linearised form of the integrated equation (2.3) :

$$\frac{t}{c} = \frac{1}{k c_\infty^2} + \frac{1}{c_\infty} t \tag{2.4}$$

The slope and intercept of the resulting linear regression determines the parameters k and c_∞. The latter can also be found experimentally.

Standard (or Characteristic) function method

The model consists of representing the extraction curve with an exponential function with three parameters A, B, H of the form (Bart and Pilz 2011; Simeonov et al. 1999):

$$c = A - B \cdot exp(-Ht) \tag{2.5}$$

where A is approximated to the concentration at equilibrium in the solvent phase. The parameters are estimated by nonlinear regression of experimental data. This approach has a good fit when the dynamics of the process approximate first-order kinetics.

2.2.2 Phenomenological modelling

The modelling of solid extraction is not as mature as other separation operations like liquid-liquid extraction or chromatography (Bart and Pilz 2011). Analytical solutions of Fickian models have been favoured to simplify the estimation of model parameters from the shape of experimental extraction curve. However, more rigorous modelling implementations have been reported with rise of supercritical fluid extraction due to the academic push into characterizing this green alternative to the conventional chemical solvents (Kassing et al. 2010b).

2.2.2.1 Derivations from Fick's Second Law of Diffusion

The desorption and diffusion steps are often considered the rate-limiting steps of the extraction, and they can be embodied into a combined effective diffusivity term. The diffusion equation is similar to the heat equation and it appeared when Fick applied this equation to the diffusion phenomena. Its application to solid-liquid extraction is a common way to estimate the effective diffusion coefficient. This parameter is an important property for equipment design as it indicates the mass transfer rate of the process.

$$\frac{\partial C}{\partial t} = \nabla(-D_{eff}\nabla C) \tag{2.6}$$

Under the assumptions of constant effective diffusion coefficient (D_{eff}), negligible external mass transfer, and perfectly mixed liquid phase in the extractor the model can be simplified and solved analytically (Crank 1975). Considering a symmetrical and porous sample particles of spherical geometry and radius R:

$$\frac{\partial C}{\partial t} = D_{eff}\left(\frac{\partial^2 C}{\partial r^2} + \frac{2}{R}\frac{\partial C}{\partial r}\right) \tag{2.7}$$

where C is the concentration of the extract compound in the solid matrix changing with time over the radius r, the initial and boundary conditions are as follow:

$$C_{t=0} = C_0, \qquad \forall r \tag{2.8}$$

where C_0 is the initial concentration of solute in the sample particle and an uniform distribution of active compounds in the solid matrix is assumed.

$$\left.\frac{\partial C}{\partial r}\right|_{r=0} = 0, \qquad t > 0 \tag{2.9}$$

indicating that diffusion flux is zero at the core of the particle due to symmetry.

$$C_{(r=R)} = 0, \qquad t > 0 \tag{2.10}$$

$C_{(r=R)}$ is the concentration of solute at the interface, which is considered to be zero at all times with the assumption of negligible external mass transfer resistance.

The differential equation can be expressed as:

$$\frac{C - C_0}{C_{(r=R)} - C_0} = 1 + \left[\frac{2R}{\pi r}\sum_{n=1}^{\infty}\frac{(-1)^n}{n} sin\left\{\frac{\pi n r}{R}\right\} exp\left\{-\frac{D_{eff} n^2 \pi^2 t}{R^2}\right\}\right] \tag{2.11}$$

subsequently the total amount of solute extracted from the sample at time t can be expressed by integrating the solute concentration over the particle radius R

$$\frac{M_t}{M_\infty} = 1 - \frac{6}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2} exp\left\{-\frac{D_{eff} n^2 \pi^2 t}{R^2}\right\} \tag{2.12}$$

where M_t is the total amount of solute extracted from particle after time t, M_∞ is the maximum amount of solute extracted after infinite time. After a certain large enough time lapse (i.e., after the washing stage) only the first term of the series in equation (2.12) is significant. The equation can also be expressed in terms of solute concentration in the extraction solvent c:

$$\frac{c_\infty - c}{c_\infty} = -\frac{6}{\pi^2} exp\left\{-\frac{D_{eff}\pi^2 t}{R^2}\right\} \tag{2.13}$$

$$ln\left(\frac{c_\infty - c}{c_\infty}\right) = -0.498 - \frac{9.87 D_{eff}}{R^2} t \tag{2.14}$$

where c_∞ is the concentration of solute in solvent after infinite time (i.e, equilibrium).

To determine D_{eff}, which is a typical goal of an experiment, a linear method can be employed by determining the slope of equation (2.14) when fitted to the experimental extraction curve.

Osburn and Katz (1944) proposed an extension of the model above to account for the two distinct regions of mass transfer, described previously in Figure 2.4, washing and the diffusion stages. This yields the following equation (for a spherical shape):

$$\frac{c_\infty - c}{c_\infty} = \frac{6}{\pi^2}\left[f_1 exp\left\{-\frac{D_1\pi^2 t}{R^2}\right\} + f_2 exp\left\{-\frac{D_2\pi^2 t}{R^2}\right\}\right] \tag{2.15}$$

where f_1 and f_2 are the extracted solute fractions for the washing and diffusion stages, with their respective effective diffusion coefficients D_1 and D_2. These parameters can also be estimated through linear regression. In the second stage of the experimental curve only the second term related to the diffusion is significant, thus the parameters D_2 and f_2 can be determined the same way as in equation 2.14. The early stage of the extraction allows for the determination of D_1 and f_1 by linear regression of the following expression:

$$ln\left(\frac{c_\infty - c}{c_\infty}\cdot\frac{\pi^2}{6} - f_2\right) = f_1 - \frac{9.87 D_1}{R^2} t \tag{2.16}$$

The application of Fick's second law can also be made for other geometrical shapes of solid other than spherical. Other analytical solutions for infinite or semi-infinite

geometries of slabs (plane sheet) and cylinders are also commonly applied (Asl and Khajenoori 2013).

These model solutions have been obtained solely by solving the mass balance transfer in the solid particles. However the mass balance in the solvent can be taken into account. The balance equation for the liquid phase is given by (Tsibranska et al. 2011):

$$\frac{dC_L}{dt} = -\frac{V_S}{V_L}\frac{d\bar{C}}{dt} \tag{2.17}$$

where $\bar{C}$ is the volume averaged solid phase concentrations at time t, V_S and V_L are the respective solid and liquid phase volumes.

Sharing the same assumptions as the mass balance in the solid particles shown previously, except for the additional initial condition :

$$C_{L,t=0} = 0 \tag{2.18}$$

and a different the boundary condition at $r = R$, respecting the negligible external mass transfer assumption:

$$C_{r=R} = C_L \tag{2.19}$$

The volume averaged solid phase concentrations $\bar{C}$ is calculated (for spherical sphapes):

$$\bar{C} = \frac{3}{R^3}\int_0^R C(r)r^2 dr \tag{2.20}$$

Wongkittipong et al. (2004) solved this model numerically, with a discretization of the particle radius with second-order finite differences using the Crack–Nicolson method. The effective diffusion is then estimated by fit with to the experimental points.

2.2.2.2 Rigorous models

These phenomenological models use classical mass transport and thermodynamic equations and they may involve the joint description of macro-scale and micro-scale aspects of the process (Both et al. 2014; Kassing et al. 2010a; Sixt et al. 2018b).

Macro-scale models

The macro-scale component is related to the mass transport in the liquid phase of a given extraction equipment and operational mode. They consist of mass balance equations of the solute in the liquid phase. The integration of these differential equations is what yield the extraction profiles, seen in Figure 2.4. Extraction profiles are thus shaped by the extraction set-up. Percolation can be described with a distributed plug flow model.

$$\begin{aligned}\frac{\partial C_L(z,t)}{\partial t} = D_{ax}\cdot\frac{\partial^2 C_L(z,t)}{\partial z^2} - \frac{u_z}{\epsilon}\frac{\partial C_L(z,t)}{\partial z} - \\ -\frac{1-\epsilon}{\epsilon}\cdot k_f \cdot a_P \cdot [C_L(z,t) - C_P(r=R,z,t)]\end{aligned} \tag{2.21}$$

where $\frac{\partial C_L(z,t)}{\partial t}$ represents the solute concentration accumulation in liquid phase. The three right-hand side terms of the partial differential equation are related to the convection, dispersion and mass transport between the solid and liquid phases.

D_{ax} is the axial distribution coefficient which determines if the flow profile deviates from that of an ideal plug flow reactor ($D_{ax} = 0$). This term can be determined through tracer experiments or through correlation between the Reynolds and Péclet number, which in turn is related to the axial diffusion through the mean particle diameter ($d_{P,mean}$), void fraction of percolated fixed bed (ϵ) and the empty tube velocity of the fluid (u_z). The last two are determined through tracer experiments (Sixt et al. 2018b). The following equations show the relationships between these parameters:

$$Re = \frac{u_z \cdot d_{P,mean} \cdot \rho_L}{\eta \cdot \epsilon}$$

$$Pe = \frac{0.2}{\epsilon} + \frac{0.011}{\epsilon}(\epsilon \cdot Re)^{0.48}$$

$$D_{ax} = \frac{d_{P,mean} \cdot u_z}{\epsilon \cdot Pe}$$

The second term in equation (2.21) is related to the convection resulting from pumping the percolating liquid. These two last terms are related to the fluid dynamics while the last term is related to the flux of the solute from the solid matrix surface ($C_P(r = R, z, t)$) to the solvent. This term is dependent on the specific surface are (a_P) and the mass transfer coefficient (k_f). For spherical particles:

$$a_p = \frac{6}{d_{P,mean}}$$

and the mass transfer coefficient (k_f) can be given through correlations of the Schmidt (Sc - ratio of convective to diffusive mass transfer), Sherwood (Sh - ratio of effective mass transfer to diffusive transport) and Reynolds numbers.

$$Sc = \frac{\eta}{\rho_L \cdot D_{12}}$$

$$Sh = 2 + 1.1 \cdot Sc^{0.33} \cdot Re^{0.6}$$

$$Sh = \frac{k_f \cdot d_{P,mean}}{D_{12}}$$

where D_{12} is the binary diffusion coefficient between the respective component and the solvent and can be determined through the following Stokes-Einstein equation

$$D_{12} = \frac{k_B \cdot T}{6\pi \cdot \eta \cdot r}$$

were k_B is the Boltzmann constant.

For an immersion operation in batch mode, the process can be assumed to be ideally stirred and equation 2.21 is simplified to:

$$\frac{\partial C_L(z,t)}{\partial t} = -k_f \cdot a_P \cdot [C_L(z,t) - C_P(r=R,z,t)] \tag{2.22}$$

Due to the broad diversity of equipments these well described models may deviate from the real system and adequate corrective modelling approaches such as defining different compartments with variant models, or defining residence-time distributions, have to be applied to approximate the model to reality (Kassing et al. 2010b).

Micro-scale models

The micro-scale model describes the extraction phenomena at the plant cell level. Thus it attempts to depict the previously mentioned steps (Figure 2.1) of an extraction: the solvent entering the solid particles, desorption of the desired components and then releasing these components into the surrounding bulk solvent. This local flux of the solute into the solvent is subsequently used as an input of the macro-scale model.

Assuming spherical shapes of the solid material the equation 2.6, from Fick's second law, can be adopted to describe the time-dependent loading of the component in the solid $\frac{\partial q(z,r,t)}{\partial t}$ along the axial and radial directions, in the following pore diffusion model (Kassing et al. 2012).

$$\frac{\partial q(z,r,t)}{\partial t} = D_{eff}\left(\frac{\partial^2 C_P(z,r,t)}{\partial r^2} + \frac{2}{R}\frac{\partial C_P(z,r,t)}{\partial r}\right) \tag{2.23}$$

where C is as before the concentration of the target component. The effective diffusion coefficient can be estimated from the experimental data.

To account for diffusion slowdown effect, due to the phenomena happening in the pores and not in a free liquid, it can be further expressed as:

$$D_{eff} = \frac{\epsilon_p \cdot \delta \cdot D_{12}}{\tau}$$

where the ϵ_p is the porosity of the solid material, δ is the constrictivity factor (diffusion speed impediment due to smaller pore sizes) and τ is the tortuosity (prolongation of the path through the spherical particle). These parameters can be determined through correlations or determined via suitable methods, such as the mercury penetration method (Sixt et al. 2018b).

The initial condition for this equation states that the solid material is maximally loaded at each location in the axial and in the radial direction:

$$q_{t=0} = q_{max}, \qquad \forall r \forall z \tag{2.24}$$

and the boundary conditions for the model consist on assuming a zero Neumann boundary for the concentration of solute at the core of the solid matrix.

$$\left.\frac{\partial C_P}{\partial r}\right|_{r=0} = 0, \qquad t > 0 \tag{2.25}$$

and considering the solute mass balance at the surface of the solid to link the micro model with the previously defined macro model

$$D_{eff}\frac{\partial C_P(z,r,t)}{\partial r} = k_f \cdot a_P \cdot [C_L(z,t) - C_P(r=R,z,t)], \qquad r = R \tag{2.26}$$

These micro-scale models can also includes a modelled representation of the adsorption/desorption equilibrium inside the pores (Kassing et al. 2010a). The chosen equilibrium curves are then linked to the micro-scale model through the loading q. Typical equilibrium isotherms are:

Henry

$$q = K_H C_P \tag{2.27}$$

Freundlich

$$q = K_F C_P^n \tag{2.28}$$

Langmuir

$$q = q_{max}\frac{K_L C_P}{1 + K_L C_P} \tag{2.29}$$

The maximum load q_{max} is the total amount of the regarded target and side components. The parameters can be estimated from measurements of the residual load and the concentration in the liquid, which yield an equilibrium curves. The are measured with multi-stage macerations (Kassing et al. 2012). The equilibrium is specific for each solvent, thus the solvent selected prior to the parameter determination should be the solvent of the desired system.

Other modelling approaches

Other approaches to the modelling of micro-scale phenomena exist and are associated to the non-homogeneous nature of the solid matrix. The shrinking core model assumes spherical particles and the phenomena follows first a desorption in the outer zone of the particle, followed by the diffusion of the inner core solute to the outer zone until it is transfer to the surrounding media. The desorption front moves from the surface to the core, hence the name *shrinking core*. The process is finished when the volume of the shrinking core particle reaches zero.

The other alternative is the broken and intact cells model which is based on the concept of target components existing in both inside the cell particles, as well as in broken and accessible vacuoles. This assumption tries to provide a phenomenological description to the observed distinct washing and diffusion steps on real extraction experiments. In the first case, there is no diffusion limitation, but in the second case there is a strong diffusion limitation of the extraction. This concept can also be implemented in the previous pore diffusion by means of describing radial pore size and solute distribution. (Bart and Pilz 2011; Sixt et al. 2018b)

2.2.3 Applications

Exhaustive reviews of model applications across the different configurations, especially supercritical fluid extraction, already exist (Chan et al. 2014; Kassing et al. 2010a; Reverchon and De Marco 2006). This subsection highlights some applications of the previous methods in conventional solvent extraction (excluding the more advanced supercritical fluid extraction applications) in stirred tanks.

A literature search for "solid liquid extraction model/modelling", after a sensible deselection of entries which are outside the scope of the work (solid-phase extraction, heavy metal adsorption from soil, etc.), reveals that the most cited research in this field are related to response surface methodology modelling. The topic is often the comparison of the newer greener technologies against the conventional solvent extraction. Temperature, solid-liquid ratio, particle size, solvent concentration and extraction time are the prevalent studied parameters in these experiments while the yields or total amounts of the desired component are the responses (Ku and Mun 2008; Prakash Maran et al. 2013a; Quanhong and Caili 2005; G. Zhang et al. 2011). Different types of experimental designs such as central composite design, Box-Behnken can be seen in the literature. Silva et al. (2007) studied the impact of temperature, time of contact and ethanol quantity in the outcomes of total phenolics, total flavanoids, and total flavonols from Inga edulis leaves. These three responses showed different extraction profiles and multi-response maximization using the simplex method (and maximum desirability for the phenolic contents) lead to an optimized set of parameters which was then validated with and external test to be within a 95% mean confidence of the predicted values. This experiments were performed in a set-up of 30 mL extractor. The majority of this types of investigations are conducted at laboratory scale, however an increasing effort has been shifted to applications at pilot scale, mostly to provide proof of concept of the newer technologies (i.e., ultrasound-assisted extraction) at larger scales (Canettieri et al. 2007; Khan et al. 2010; Pingret et al. 2012; Sorsamäki et al. 2010; Widyaningsih et al. 2018). However studies at industrial scale are scarce (Rostami et al. 2014), comprehensibly so due to the necessary experimental expenditure and excitation of operational parameters being prohibitive in a normal production environment.

Phenomenological modelling is predominantly of the Fickian type (Chan et al. 2014). The modelling in these studies serve purpose of assessing the influence of

the operating parameters (i.e., temperature, nature of the solvent and particles size) or different operation modes (i.e, conventional vs ultrasound) in the kinetics of the extraction by means of estimating the diffusivity and mass transfer coefficients in the model and thus determining which condition is more favourable (Cacace and Mazza 2003; Chan et al. 2014; Ji et al. 2006; Wongkittipong et al. 2004). Ultimately, these authors interpret their results and point to optimized operations or designs based on the outcomes of their estimated parameters. However, most times this is not achieved in a systematic way. Moreover these models are solely utilized for kinetic studies to assist in design and optimization purposes. An outlier is the work of Simeonov and Nedialcova (2006) where a characteristic function model is used to characterize the kinetic curves of three immersion stages. This model is then theoretically adapted to mimic a continuous counter-current screw reactor and the reactor length necessary to reach full equilibrium in the first stage is calculated. The model can also integrate screw revolution, which at constant screw dimensions (defined by the length calculated for the first stage), provides a manipulated variable to ensure the process reaches complete leaching.

2.3 Model-based systematic approaches

Professor Jochen Strube's group at the Clausthal University of Technology (Germany) has been developing over the past years a structured approach to process development of plant-based extractions. In the following sections these will be reviewed:

2.3.1 Applications in process development and design

In a first iteration, Kassing et al. (2010a) suggests an approach to systematically develop extraction processes, and provides a standard lab set-up to achieve this goal. First, an assessment of which solvent has the highest solubility for a given target solute is made through analysis of the solubility constant, through Hildebrand one-parameter model, $\delta = \sqrt{\frac{\Delta H_v - RT}{V_m}}$, or by assessment of the activity coefficient. Successively, the authors propose a combination of design of experiments and rigorous modelling that incorporate botanical aspects on a cell scale. The statistical models from the DoE serve the purpose of identifying the significant process parameters. Additional experiments are then needed to determine the model parameters for rigorous modelling. These experiments would be conducted on the suggested standard apparatus which comprises of maceration, percolation and Soxhlet extraction coupled with several analytical equipment such as UV-diode array detectors, evaporative light scattering detectors and a refractive index detector. These serve the purpose of quantifying the solute's concentration in the solvent. The proposed set-up finishes with further fractioning and analysis with coupling to high performance liquid chromatography and gas chromatography with the possibility of mass spectroscopy and nuclear magnetic resonance. A validation of the resulting models is made at a mini-plant,

and the optimal process conditions is identified by setting a target function on yield and purity or specific cost of goods. The same authors in a later publication provides two case studies, for both vanilla beans and pepper berries, to illustrate the proposed approach (Kassing et al. 2012). Both DoE based modelling and rigorous models are analysed. From the DoE-based analysis, insights are taken on what are the most influential parameters for equilibrium (i.e., solvent-feed ratio) and extraction kinetics (i.e., temperature, mean particle size) and what are the optimal operational parameter combinations. The rigorous models were supported by botanical investigations on pore size, porosity and compound distributions assessed through microscopy and vibrational spectroscopy techniques. The macro-model parameters such as the axial dispersion coefficient, bed and pore porosity with tracer experiments on the percolation set-up.

The total amount compound in the raw material and the desorption isotherm were then characterized via multi-maceration experiments at different solvent-feed ratios and found to be of Langmuir type for the target compound of pepper berries ($K_H = 1$) and linear for the vanilla compound. The effective diffusion parameter was then fitted to the data of a percolation experiment with short residence time. For both raw materials, the model exhibits a great fit in the first bed-volume, but a slight over-estimation at the later periods of the percolation extraction. The models were validated against other percolation experiments, at much longer residence times (lower particles sizes and flows). From this validation the authors identify the mismatch between the data and the model prediction and suggest that this is due the fact that the estimated diffusion in the previous experiment is capturing phenomena from a convection (from the particle fluid film to extraction bulk) limited operation, while the validation experiments are pore diffusion limited. A lower effective diffusion (which would yield parallel to last points of the extraction) is proposed instead of the previously fitted parameter. This yields a better fit in the validated experiment, but far worse in the original low residence time experiment to which the author suggest is due to a lack of particle size gradient in the model which would better convey the phenomena at the different stages of the extraction. The statistical model and the rigorous model for pepper are then compared for their predictive quality. Both models exhibit good prediction accuracies $89 - 100\%$, with the rigorous model exhibiting better performance, the exception being the low residence time experiment ($\sim 65\%$) as stated above. These techniques are not actually incorporated, but used together to obtain process understanding and interpret where the process optimization could be made. In this work the optimizations are done rather empirically after the analysing previous experiments. For the vanilla extraction, solvent composition is varied and an optimization from the original 100 vol.% ethanol to a 50/50 vol.%, leads to a 10-15% increase in total cumulated target compound but with slightly slower kinetics. For pepper berries, other alternative techniques are implemented and compared with the benchmark percolation and macerations. However, no improvement from the benchmark percolation mode is achieved. This method is systematic in the sense of the model development and understanding of the process follows a guideline for experimentation and model development. Nevertheless, the optimization approach

presented remains to be empirical, relying on the expertise of the user. The optimization scenarios are proposed based on the insights gathered, and posteriorly to new experimentations with these settings, and optimal process setting is chosen in comparison to the benchmark process used for model development.

Both et al. (2014) further expands this approach with a more detailed focus on the integration of botanic and thermodynamic aspects, with the main motivation of selecting optimum extraction equipment through phenomenological models. A model-based comparison with validated models is suggested, such that several extraction scenarios with different equipment and operating conditions (i.e, percolation with different flow rates, (multi-stage) maceration, recycling loop, counter-current percolation) are tested on the basis of measures such as yield, purity, space-time-yield, equipment/apparatus efforts, solvent consumption, extraction time and solid-liquid ratio. The authors propose a spider chart as graphical tool to assess how each set-up rates among each other. This approach becomes truly model-based for process design.

2.3.2 Applications in process optimization

Besides process design, the guidelines and tools provided can be used to optimize established fullscale processes as it is shown by Both et al. (2013) for the example of sugar extraction from sugar beets. Here the authors employ the DoE statistical modelling to obtain the optimal operating conditions ($65 - 70°C$ and pH=6.5) and consequently execute the experiments as proposed by Kassing et al. (2010a) at these settings, in which the rigorous model parameters are estimated. The fluid dynamic parameters (i.e., axial dispersion, external porosity) are determined at laboratory scale and industrial scale. The model is used for validation of process data and the rigorous model identifies that with a smaller external porosity the same efficiency of extraction can be achieved with less volume, thus enhancing volumetric productivity.

2.3.3 Applications in process monitoring and control

The combination of at-line analytics for raw material analysis has also been subject of study in the resource efficiency of natural product extractions (Gudi et al. 2015). Sixt et al. (2018a) implements the rigorous modelling systematic approach in the context of process control, combining it with an in-line Raman spectroscopy tool for the determination of the target compounds anethole and fenchone in real-time during the extraction of fennel with ethanol. The PAT tool development is made through partial least squares modelling and achieves high accuracy performance for both compounds. The data is directly linked to the rigorous model and predictive simulations are performed while the data from the Raman spectroscopy is processing.

2.4 Pectin extraction

The industrial production of pectin has been around since the early 20^{th} century when the preserves industries acknowledged its gelling power. These manufacturers produced a pectin-rich extract from the fruit trimmings leftover from the fruits which were used in jam making. This extract would later be added to the jam making process to supplement the pectins extracted from the fruits and ensure a satisfactory gelling. The size of the juice and cider industries, and the perspective of valorization of waste through a speciality by-product led to the creation of dedicated pectin producing enterprises. In this section, a review of the current industrial processes, potential alternatives as well as efforts made in modelling this process are presented. A general overview of the pectin plant set-up at CP Kelco will be highlighted.

2.4.1 Commercial extraction process

Nowadays, all pectin production in the industry comes from its extraction from plant material. The production can be crudely separated into raw material pre-treatment, extraction and post-extraction stages. Pectin is widely available in across different fruits and plants in nature. However, industrial production is mostly confined to the conversion of citrus peels (≈85%) and to lesser degree apple pomace (≈14%) due to their prevalence in the juice industry, which consequently leads to large waste quantities of these fruits (Ciriminna et al. 2015). Another source of pectin which has been implemented at an industrial scale is sugar beet pulp, but the gelling properties of the pectin extracted limits its applications and therefore its production potential.

Nevertheless, pectin from different sources has been extracted and tested at a laboratory level as tomato, carrot, and many other fruits of high consumption and waste generating volume such as banana, watermelon and pumpkin (Dranca and Oroian 2018; Maric et al. 2018). Different plant materials have different quantities of pectin, which will in turn have different physicochemical properties than pectins extracted from other sources. Pectins derived from apple pomace produce more viscous gels, however suffer from a brown discolouration which has negative impact in several food applications. In turn, citrus peel have higher quantity and these pectins have a lighter colouring. However, as Dranca and Oroian (2018) highlight, the origin (geographical) will greatly affect the pectin content and total content of similar waste-streams. Even within the same suppliers, variation is unavoidable due to different seasons and treatments given to that specific batch.

Pre-treatment is necessary to increase stability of the raw material which is a waste-stream of another industry. Pectin quality is affected by the physical and chemical operations in the preceding supplier (Sakai et al. 1993). Pectin can be extracted from wet-state raw material, however the process is conditioned to the existence of nearby suppliers. For storage and transportation purposes, peels and pomace needs to undergo a drying, washing or blanching step. The material stemming from its native industry is in wet-state and is susceptible to fungal contamination. These molds produce pectin methylesterase, polygalacturonase and other pectin lyases

which degrade and de-esterify the pectin in the raw materials. Storing wet pomace or citrus peel even for the time needed for transportation from the juice plants to a central location can induce a loss in quality of the extracted pectin. Therefore, the raw material is typically washed and dried at temperatures sufficient to destroy the enzymes and funghi, but without sacrificing too much in pectin content (May 1990).

The process flow of pectin plant is described in Figure 2.5. The production process, post raw material handling, entails the following core steps: Solid-liquid extraction, purification, precipitation, drying and milling. An optional step of further de-esterification or amidation can be added, prior to drying and milling, depending on the desired product specification. The typical extraction is carried out with a certain peel to hot acidic solvent ratio. The downstream process follows a sequence of purification steps to ensure a supernatant with high clarity, free from peel and cellulose residues. This pectin rich juice is concentrated prior to precipitation with alcohol. Lastly, the precipitate is dried and milled to what is called as semi-finished product. Commercial pectins are standardized products, in order to ensure gel strength consistency in their final applications. Therefore, the semi-finished product is blended with other semi-finished from other batches, as well as with sugars, according to the order set in place to attend to the customers gelling and thickening requirements (Sakai et al. 1993).

The chemical extraction with mineral acids (i.e., hydrochloric acid, sulphuric acid, nitric acid) is the conventional way of producing pectin, due to its price and satisfactory production yield. Other organic acids have also been used (i.e., citric acid, acetic acid) in countries where regulations against the use of mineral acids in foodstuff production are enforced. Under identical extraction conditions, citric acid has been reported to have the best extractive performance (Dranca and Oroian 2018). In addition, the extraction success depends on numerous other factors such as temperature, pH value, solid to solvent ratio, raw material particle size and other properties which affect the diffusion rate (Maric et al. 2018).

The conventional process is ran at temperatures between T=60-100°C and at pH of 1.3-3 (Putnik et al. 2017). The solid-liquid ratio (S/L) is also an important factor

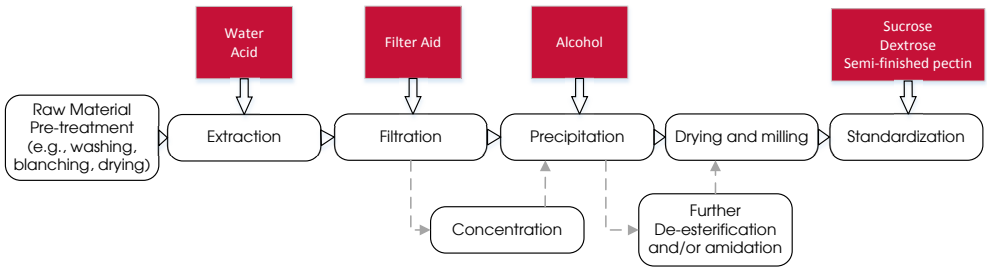

Figure 2.5: Simplified scheme of the pectin production process constitutive steps. Dashed arrows symbolize optional routes.

as the pectin extract becomes highly viscous as the extraction progresses, affecting negatively the mass transfer rate. A higher (S/L) provides a better concentration gradient and less viscous broth, leading to a higher extraction rate and yield. However, economical feasibility has to determine the optimized use of solvent, and this ratio is normally around 1:18 (w/v) (Sakai et al. 1993). This parameter has to optimized depending on the system, for example Kulkarni and Vijayanand (2010) have shown that, for a particular passion fruit peel, no improvement above 1:30 (w/v) is observed. Smaller particle size leads to increased surface contact area and to smaller diffusion path for pectin. However, it is also necessary to have in consideration the efficient downstream processing of the resulting slurry. This limits how small the particles can be and consequently how fast the diffusion kinetics are, with respect to the pore size.

Extraction is a key process unit, in the sense that the operation conditions lead to a certain pectin quality which affects the remainder of the process line operation if a certain quality target profile is required. This quality profile comprises certain critical quality attributes, most importantly %DE for standardization purposes. Pectin types fall under three denominations: High methoxyl pectins (HM), low methoxyl pectins (LM) and amidated pectins. HM pectins have $\%DE > 50\%$, form gels in high sugar concentration and low pH applications and are not reactive with calcium ions. At an extraction level, HM pectin offers more flexibility for further downstream quality manipulation. In contrast, LM pectins may form gels at a broader pH range and lower sugar concentrations in the presence of calcium cations (Ca^{2+}). Amidated pectin requires de-esterification with amonia (typically to LM pectin levels), leading to the substitution of ester groups with amide groups. This product has different gelling properties than the prior two pectin types, needing less calcium to gel and show more stability at higher pH values (Ciriminna et al. 2015; May 1990). Despite a growing interest in the latter type, the conventional pectins still provide a big market share, as some product makers still prefer to avoid the declaration of amidated pectin in their label. In addition from the %DE and degree of amidation ($\%DA$) specifications, another quality attribute that can be considered of interest is intrinsic viscosity (IV), which is linked to the pectins molecular weight which decreases when the homogalacturonan chain is broken up. The IV is also affected by the composition and shape of each component in the polymer (Cho and Hwang 2000).

2.4.2 Modelling

In the literature, the amount of response surface methodology approaches far outweigh the studies which investigate mathematical models of the extraction of pectin. The most cited response surface methodology articles can be found in Table 2.1. It is visible that a focus on the study of the effects of temperature, pH and time are prevalent. There is also a focus in studying the extraction from a variety of raw material, from banana and sweet potato peels (Bee Lin and Yek Cze 2018; Oliveira et al. 2016; Y.-Y. Zhang et al. 2013) to the more conventional citrus and apple peels (Sathish et al. 2016; Sebaoui et al. 2017; X. Wang and Lü 2014).

Table 2.1: Most cited response surface methodology applied to pectin extraction papers.

Authors & citations	Extraction	Factors	Response	Significant factors	Outcomes (Optimal conditions)
S. Wang et al. (2007) 179	Microwave assisted extraction	Extraction time pH (HCl solution) S:L ratio Microwave power	Yield	pH Mw power pH and S:L ratio interaction	pH=1.01 mw power=499.4 W time=20.8 min solid:liquid ratio=0.069 pectin yield=15.75 g/g%
Canteri-Schemin et al. (2005) 112	Conventional	Citric acid concentration Extraction time	Yield	linear and quadratic effects of both variables	Time=152.85 min citric acid=6.2 g/g% pectin yield=17.82 g/g%
Pinheiro et al. (2008) 105	Conventional	Citric acid concentration Extraction time	Degree of esterification	Citric acid concentration as most important factor	Citric acid =0.086% w/v Time=60 min %DE=78.59%
Masmoudi et al. (2008) 102	Conventional	Temperature Extraction time pH	Yield	Temperature as most important factor	Optimal conditions: T=84.34°C pH=2.8 time=3h34min pectin yield=11.21%
Prakash Maran et al. (2013b) 81	Microwave assisted extraction	Microwave power Irradiation time pH Solid-liquid ratio	Yield	All	Microwave power = 422 W irradiation time = 169 s pH =1.4 s-l ratio=1:16.9 pectin yield=19.24%
Gan and Latiff (2011) 58	Conventional	Temperature Extraction time pH	Yield Uronic acid Antioxidant activity	Yield and Antioxidant activity include quadratic with interactions. Uronic acid is linear	T=(1) 80 (2) 67.7°C time=(1) 3.93 (2) 3.67 h pH=(1) 2.45 (2) 2 pectin yield=12.1% uronic acid =20.6 mg/g Antioxidant activity (1)251.7 (2)225.1%
Zykwinska et al. (2006) 56	Conventional	NaOH concentration Temperature	Composition (neutral and acidic sugars)	NaOH less significant than temperature	Analysis of the arabinan and galactan chains of pectin extracted at different conditions
Kliemann et al. (2009) 54	Convetional	Acid type Temperature pH Extraction Time	Yield	Temperature pH time	Citric acid T=80°C pH=1 time=10 min pectin yield=70% (extrapolation)

The use of response surface methodology is also favoured for the analysis of alternative extraction techniques and is a very present topic in the literature, with enzyme-assisted (Sabater et al. 2018), microwave-assisted (Prakash Maran et al. 2013b; Seixas et al. 2014; S. Wang et al. 2007) and ultrasound-assisted (Minjares-Fuentes et al. 2014; W. Wang et al. 2015). However, a common denominator in all these studies, to the best of this author's knowledge, is that they have all been performed at laboratory scale, and thus its applicability for other set-ups at bigger scales is limited.

A few kinetic empirical modelling studies of the pectin extraction exist and exhibit various degrees of success in describing the extraction kinetic curves, with Peleg's model showing particular adequacy in the case of pectin extraction from red dragon fruit peels ($R^2 = 0.96$) (Chua et al. 2018).

Some phenomenological modelling approaches of the pectin solid-liquid extraction have been reported. The first attempt can be attributed to Panchev et al. (1989), who proposed a simple kinetic model taking into account the hydrolysis of protopectin to pectin and its degradation:

$$\frac{dC_{protopectin}}{dt} = -K_{hydro} \cdot C_{protopectin} \tag{2.30}$$

$$\frac{dC_{pectin}}{dt} = K_{hydro} \cdot C_{protopectin} - K_{deg} \cdot C_{pectin} \tag{2.31}$$

where K_{hydro} and K_{deg} are the rate constant of the transformation of protopectin into soluble pectin and the rate constant of pectin degradation. The model parameters were estimated with data from an apple pomace extractions with 0.5% nitric acid at temperatures of 60°C, 70°C and 80°C for a period of 120 min (sampling every 10 min) with a good fit. The approach of modelling the system as a consecutive reaction mechanism proves to work in this particular system as the second rate constant presents the curve features of a process with internal and external diffusion. Pagán and Ibarz (1999) used the same rationale and obtained the same satisfactory results and have studied the dynamic influence of temperature and pH on extraction and degradation kinetics.

Cho and Hwang (2000) were the first to model quality parameters of the dissolved pectins. The authors also followed the consecutive reaction mechanism for yield and modelled the intrinsic viscosity with molecular-dimension-to-viscosity conversion factor to form a first order-reaction, using an exponential form with temperature for this parameter. This article was also the first instance where the concentration reaction rates were expressed as Arrhenius-type equations with respect to temperature. The models showed good levels of agreement with the data ($R^2_C = 0.922$ and $R^2_{IV} = 0.682$). An extension of this model was used to study the kinectics when ultrasound and/or heating are applied (Xu et al. 2014). Sebaoui et al. (2017) also adopt the Panchev et al. (1989) model to study the kinetics of a lemon peel extraction, whilst using data from a response surface methodology model to estimate the kinetic parameters.

A more comprehensive model, attempted to describe the process as a first order chemical reaction in the solid phase, including internal diffusion of pectin in the solid pores and external mass transfer at the solid boundary layer (Minkov et al.

1996). The model detailed the transport of pectin in the peel (internal diffusion) through a partial differential equation, and showed good fit to the data. More recent implementations with this level of description are given by Durán et al. (2015) and Andersen et al. (2017) and include the degradation effect on solubilized pectin. The former is implemented on the extraction from mango peels and it resorts a lattice Boltzmann method for the numerical simulation of the model, whilst the latter is applied to citrus peels and uses the method of lines for the discretization of the partial derivative equation in the one-dimensional peel slab. Furthemore, the model developed by Andersen et al. (2017) relates process parameters (i.e., temperature, pH and raw material attributes) with the critial quality attributes $\%DE$ and IV. This model will be used in this thesis as basis and will be further detailed in section 5.1.2.

2.4.3 Pectin extraction operation at CP Kelco, Lille Skensved

The general pectin extraction process at CP Kelco is represented in Figure 2.6. The extraction set-up is composed as battery of batch tanks, which operate in parallel. Dried peel, in coarsely ground form, is fed to the process in controlled weight according to the targeted recipe. Once the mixing tanks have been filled with the solvent mixture, and reached the desired extraction temperature, the peel is added to the tanks. The mixture is carefully agitated and recirculated to ensure thorough mixing. The mixture is pumped to extractor tank where no further changes can be made to the process conditions. Therefore, once the actual extraction operation begins it will progress at the pre-determined conditions.

Currently, operational conditions setting for pectin extraction are based on the peel type that is to be deployed in production. This is dealt in a heuristic fashion. When handling a known peel type, the extraction conditions are inferred from historical data, ideally if the product previously has been manufactured using this peel. If not, a possibly lengthy trial-and-error procedure is carried out to determine a set of feasible conditions. The quality measurements such as pectin $\%DE$ and IV are not routinely checked during a batch. This quality feedback on the extractions performance is only performed further downstream as well as at the end of the production line. This may lead to an adjustment of the temperature and/or amount of acid

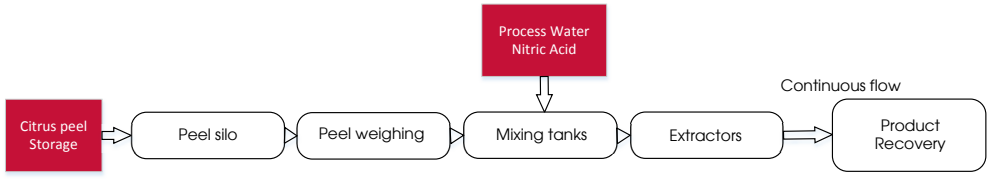

Figure 2.6: Simplified scheme of the pectin extraction process constitutive steps and their respective number of units.

added in the next extraction batch that has the same peel-product combination. If the incoming peel is of unknown performance, the production reverts to information from the same country of origin, supplier, fruit type, etc. and it can take over a day for the semi-finished pectin product from a given peel to be analysed in the lab. This ultimately leads to product produced at suboptimal conditions. For some products the IV variable is seen as a very desirable attribute to "maintain as high as possible". However, there are no explicit tools that can drive its optimization and only sub-optimal heuristic actions are currently taken.

2.4.4 Datasets for case development

Different sources of data are analysed in this thesis with the ultimate goal of developing the tools necessary for the proposed operational strategy in the following chapter. The three different datasets types can be separated by the scale of the equipment and are summarized in Table 2.2. Experiments performed in laboratory scale are used as the reference variables in the raw material assessment approach comparison. These are coupled with Fourier-transformed near-infra red spectra for each peel sample. The dataset developed by Andersen et al. (2017) at pilot scale, where one type of lime peel is extracted in a 3 full factorial scheme plus 3 water extractions (no addition of NO_3) at different temperatures, is used to evaluated not only the derived model and the identifiability of its parameter estimation, but also to explore other types of models including empirical models and shed light on the relationship between the critical process parameters and the model output. In addition to the lime peel pilot scale design depicted in the table, four water extractions were ran for lemon (3 extractions) and orange peel (1 extraction) in similar temperatures ranges. These are used to showcase the modelling mismatch to data provenient from different raw material than the one used in the parameter identification, even though the system and conditions are the same. The full-scale production data are be used to assess if the model predictions are enhanced given in-process measurements.

Table 2.2: Summary of the different types of datasets used in this work to develop and assess the necessary tools for the proposed operational strategy. The system specifies the physical hardware in which the experiments were generated is specified as well as the types of samples(raw materials) used, conditions and measured outcomes. In the applications column the main result and its respectiove section is referenced.

Scale	System	Description	Application
Laboratory	Flask immersion	• 84 different peel samples Lemon (42), Lime (26) and Orange (15) Standardized conditions: $T = 70\,^{\circ}\text{C}$, 3 g peel, 150 mL 49 mM nitric acid Samples at 20 and 240 min for Concentration, $\%DE$ and IV	Development and comparison of approaches for assessment of raw materials in **Section 4.3**
Pilot	DTU immersion	• Lime peel Full-factorial DoE T:{60 °C, 70 °C, 80 °C}, pH:{1.5, 2.3, 3.1} 3 water (no-added acid) extractions Extraction profile data for Concentration, $\%DE$ and IV Data generated by Andersen et al. (2017)	Empirical and model-based analysis of the Table 5.1 model in **Sections 5.2 and 5.3**
	CP Kelco immersion	• 2176 different peels Conditions mimicking fullscale temperature, pH and s:l ratio End-product related data	Assement of raw material impact in historic performance in **Section 4.2**
Production	Stirred vessel immersion	• Lime peel 12 extractions in NOC* range Extraction profile data for Concentration, $\%DE$ and IV	Assessment of in-process improving predictor for batch quality forecast in **Section 7.3**
		• Lime peel 2 extractions in NOC range Extraction profile data for Concentration, $\%DE$ and IV	Used in showcasing the integrated operational strategy in **Chapter 8**

* NOC - Normal Operating Conditions. Comprises perturbations in temperature, pH and peel amount that are expected in production behaviour.

2.5 Summary

This chapter reviewed the base concepts as well as the existent developments in process systems engineering within solid-liquid extraction. The current practice specific to pectin production was highlighted, and the particular operation at CP Kelco was introduced. The data used to develop this case study was also summarized. Substantial modelling efforts of different kinds have been made throughout the years, and despite the majority of the applications still refer to empirical models, in the last decade more mechanistic approaches have been brought forward and incorporated into different applications. Some of these take inspiration from chromatography modelling and have a high degree of sophistication. However, within all the different applications there was still opportunity to devise an operational strategy which promotes ongoing process verification and improvement. This is introduced in the next chapter. In terms of the pectin extraction case study, the operation relies on the heuristic knowledge acquired throughout the years. Therefore it would benefit from the development of tools that assess raw material quality and link that information with process parameters and product quality attributes. The application of the proposed operational strategy to the pectin extraction case would provide CP Kelco with more informed decision-making and a concerted optimization of the process conditions for a given peel and desired pectin quality profile.

CHAPTER 3

An Improved Operational Scheme

This chapter proposes a knowledge-based operational strategy for processes where raw material quality fluctuation is a critical concern. Section 3.1 is dedicated to analysing the problem at hand in the quality-by-design (QbD) frame and furthermore provide a structured operational strategy. Section 3.2 highlights the steps in which there is a need for process analytical technology (PAT) and model-based procedures.

3.1 A QbD-based process operation

Keeping a constant quality target product profile on production of natural products from feedstock with natural variability is a critical challenge. Traditional production processes for these operations are initially designed and validated with heuristic knowledge from few production lots. When established, these processes flexibility for rapid changes to improve product quality or efficiency of the process is severely constrained. The quality control relies on exhaustive laboratory, often with poor automation and data usefulness (Uhlenbrock et al. 2018). In a competitive industry where technology is not a differentiating factor, the production efficiency and ability to adapt to the customer needs are key drivers for success. A sequenced approach, beginning with the clear definition of the customers needs translated into product attributes and then into unit operation processes targets and parameters is highly beneficial. This has to come as a highly collaborative cross-functional effort in a production organization to be able to identify and address the critical aspects regarding processes and products.

There is a need to be adaptable and have a comprehensive understanding of process unit operations, which together with process analytical technologies, are true enablers of continuous improvement as it is stated in the guideline Q8(R2):

"Product and process understanding in combination with quality risk management will support the control of the process such that the variability can be compensated for in an adaptable manner to deliver consistent product quality"
(ICH Expert Working Group (2009))

Building quality into a product by the way the process is developed and operated is the fundamental idea behind QbD rationale. This consists on establishing scientific and reproducible methods that improve the understanding of relationships between process parameters and product quality attributes, which can be used to support an operational strategy that undergoes continuous process verification/improvement during routine manufacturing. Even if product (or semi-finished product) testing is a necessary part of quality control and compliance, production quality can be anticipated if there is an understanding of the relationships between process settings and material attributes as represented in Figure 3.1. This may enable a tailored production as well as anticipate quality concerns allowing for their impact to be mitigated.

For a QbD implementation, a thorough understanding of the underlying relationship between materials, process parameters and final product attributes is necessary (Rathore and Winkle 2009). A QbD implementation involves the steps described in Figure 3.2:

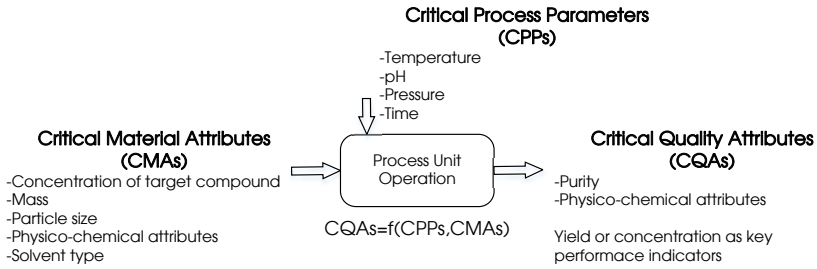

Figure 3.1: Link between between critical material attributes, critical process parameters and critical quality attributes. Examples of each are shown.

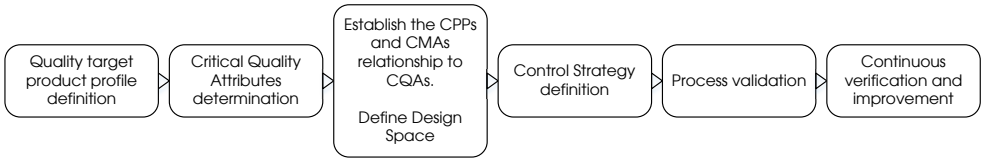

Figure 3.2: Sequence of the Quality by Design approach to product development and manufacturing

- Identification of quality target product profile and critical quality attributes (CQAs) essential for the right product application and safety;
- Process design in order to deliver product specifications, this includes the identification of the critical process parameters (CPPs) and how they interact;
- Development of a quality control strategy to guarantee consistent process performance to the required CQAs;
- Ongoing process verification efforts to ensure the process reproducibility over the lifecycle of the product.

3.2 An improved operational strategy

In productions where multiple products (or multiple qualities of the same product) are derived from the same raw materials in different production series, the CQA target will not be a static one but will instead change. In order to fit to the desired product target profile, the process parameters should be adapted (within a defined design space) to cope with this change. A consideration to have in mind is that an optimal production is not found at individual optima of the unit operations which compose said production. A plant-wide approach to integrate and optimize the production is necessary, after which the individual unit operation goals are defined. This however outside the scope of this thesis.

If the combination of the target CQAs with the selected raw materials CMAs demands a requirement of critical process parameters outside the design space (e.g., due to equipment limitations, safety issues, economical feasibility), then it is necessary to re-define the process. This can be carried out either through a new CQA target which is achievable for the given raw material, or by selecting a different available raw material lot which achieves the desired CQAs, based on the prediction made by the developed tools. This is showcased in the "Series definition" stage of Figure 3.3.

As the process runs with optimized process conditions, which were given by a mathematical depiction of the system, it necessary to assess the deviation of the model from CQA measurements of the ongoing process (Figure 3.3 "Forecast correction"). Ultimately this is required since our model representation of the system may suffer from unaccounted discrepancies from the real plant (or unmodelled effects) and these have to minimized in order to operate the process, based on model optimized settings, within an acceptable degree of trust. This deviation may also result from poor raw material attributes definition, which lead to over/under estimation in the optimization routine. If such is the case the model needs to be corrected in order to be reconciled with the actual data from the process. At this point, our trust in the model-based tool is increased and the objective to which we have proposed has not changed; thus it is necessary to determine if the optimal conditions set initially are still acceptable, or if the actual production quality trajectory is wrong (Figure 3.3 "Process conditions correction").

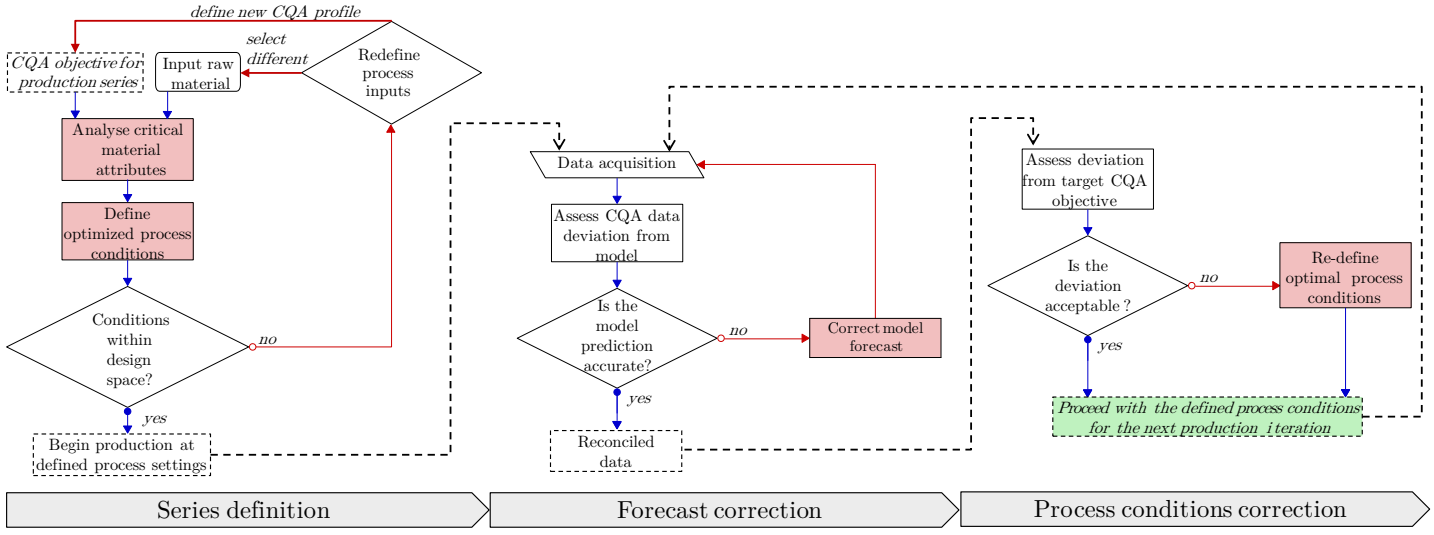

Figure 3.3: Full QbD-based operational strategy. The dashed arrows highlight the interconnections between stages. The red coloured boxes highlight steps which are enabled by PAT and model-based tools

If the latter is the case, then it is needed to re-iterate the optimization routine (now with a corrected model) and set the critical process parameters to the newly defined optimum. This procedure can be continued, in a mixed feed-forward and feedback mode, such that raw material quality can be identified prior to the production series and the mismatch can be corrected during the production series.

3.2.1 PAT tools to transition from fixed recipe to knowledge-driven manufacturing

Tools which enable the change towards a QbD-approach include prior knowledge on product and process, risk assessment methodologies, design of experiments (DoE), mechanistic models, and process analytical technology (PAT) (Yu et al. 2014). As highlighted by Figure 3.3, the operational strategy relies on some of these tools to be ready to impact production in a fast manner.

The first step which requires these tools is the raw material characterization stage. Current traditional bioprocessing plant manufacturing does not tackle raw material variability with enough operational flexibility, resulting in product quality variability. Unlike other process conditions, the manufacturer does not directly control raw materials. The quality is highly dependent on external vendors. Furthermore, raw materials may vary from lot to lot on a long timescale. In these cases, the measurement of critical properties of the raw material can allow for dynamic monitoring and control (Lanan 2008; Rathore et al. 2014). The need to tackle the natural variability of raw materials is considered an important challenge in industrial biotech processes, and an adequate approach to characterization can mitigate production performance issues and undesired deviation of the critical quality attributes (Mevik et al. 2004; Rathore et al. 2014; Skibsted 2011). Identification and measurement of key raw material characteristics (physicochemical or (micro)biological properties) through conventional analytical chemistry are an immediate first step into working towards this paradigm. Statistical analysis and modelling of this data can be performed to further enhance the use of raw material in process and learn to better classify different lots of the same raw materials (Berget and Næs 2002b). However, the measurement of raw material quality may be prohibitively time-consuming for in-production monitoring and optimization applications (Jørgensen and Næs 2004). This limitation can be overcome by coupled use of advanced spectroscopic methods such as near-infrared (NIRS), UV-visible and Raman spectroscopy with chemometric techniques. This combination constitutes a process analytical technology (PAT) tool. The PAT initiative has been defined by the US Food and Drugs Agency as

> *"(...) a system for designing, analysing and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality"*
>
> (Food and Drug Administration (2004))

In this context, PAT has been implemented in fingerprinting and classification of raw materials (Jose et al. 2011; Kirdar et al. 2009; Rathore et al. 2014).

The challenge of producing optimally for different CQA targets and raw materials can be tackled by establishing the underlying relationship between the materials, process parameters and final product attributes in terms of a mathematical model. In literature, most models which attempt to encompass empirical in their nature (polynomial, etc.) and rely on designed experiments for the modelling of process variables and then incorporate raw material variation. Their extensive use is symptomatic of the difficulty in obtaining mechanistic models that express the true relationship among process variables, raw material variables and end-product quality. If mechanistic models are available, they should be used as it provides greater level of understanding of the unit operations and processes. This is stated by multiple authors throughout the literature (Berget and Næs 2002a; K. V. Gernaey et al. 2009; Lencastre Fernandes et al. 2013; Rathore et al. 2014). Depending on the depth of the model, it can improve our process understanding in the sense of providing extended manufacturing knowledge, allowing for questions such as *"what types of raw material properties or process parameters have a higher sensitivity for end product variability?"* and *"Which are limit processing conditions, or threshold CQA targets, for a given type/class of raw material?"*, to be addressed.

A model-based optimization strategy design requires the identification of critical process parameters and definition of the objective function. The different optimization scenarios are defined according to quality specification of the particular product variety to be produced. The critical process parameters can then optimized for each scenario in a deterministic manner, or alternatively, in a robust optimization which determines the optimal operating conditions in consideration of the feedstock uncertainty and the required characteristics of desired product. A model that has been deemed adequate for the representation of system may experience mismatches for different production series, due to model parameter alterations inherent to a particular raw material which go un-modelled. Thus a continuous verification and improvement is necessary to complement the tools that have been developed.

In this thesis the development of the necessary tools for the highlighted steps will be showcased individually and then illustrated in combination (see Figure 3.4).

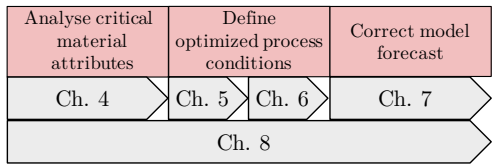

Figure 3.4: Thesis chapters of tools development for the operational strategy, applied to pectin extraction.

Part II
Development of Tools for an Improved Operational Scheme

CHAPTER 4

Raw Material Characterization

The goal of this chapter is to showcase the impact of raw material variation in pectin production and to develop the critical material attribute assessment approaches necessary for the strategy proposed in Chapter 3. Section 4.1 describes the materials and methods used both in the analysis of the raw material impact on end-product quality throughout the years, and in the development of multivariate methods to assess critical material attributes. Section 4.2 presents the raw material impact in pectin production through the use of a historical dataset. Section 4.3 presents a quantitative and qualitative comparison of three different raw material assessment approaches. Two are based on the historical statistical determination of expert-knowledge defined or unsupervised clustering classes, whilst the third relies on PAT-based calibrations for the prediction of quality parameters in the raw material from near infra-red spectra. The approaches are performed in the same dataset context.[†]

[†]This chapter includes sections which are based on the published article:
Caroço, R.F., Bevilacqua, M., Armagan, I., Santacoloma, P., Abildskov, J., Skov, T., Huusom, J.K. (2018). "Raw material quality assessment approaches comparison in pectin production". In: *Biotechnology Progress*

4.1 Materials and methods

The following material resources were utilized for the purpose of assessing the impact of different raw materials in the pectin end-product characteristics (Chapter 4.2) and in light of performing a comparative analysis of different critical material attributes assessment approaches (Chapter 4.3). The reference analytical methods are in-house standard procedures while the spectroscopic and chemometric methods are chosen considering their adequacy and widespread presence in the industry.

4.1.1 Dataset for evaluation of peel impact on product quality

The analysis in this section utilizes data from CP Kelco. This involves a *standard extraction* (standardized temperature, added acid and time) at pilot-scale, and registered information of the peels origin (country and supplier), year and the properties of its end-product pectin.

The variables used in the analysis are related to the end-product properties (i.e., IV, $\%DE$, Whiteness (L-value)) or related to end-product application (i.e., gelling strength (SAG-method)). This information is available for a period of 11 years of peels used in production. A total of 2176 samples were considered for the analysis.

4.1.2 Dataset for material assessment approach comparison

Citrus peel samples

A total of 85 raw material samples are considered, with lemon being the most represented fruit (43 samples), followed by lime (27 samples) and, orange (15 samples). The citrus peel material was ground, and no further pre-treatment of the samples was performed. CP Kelco kindly provided all sample material.

Chemical reference

CP Kelco provided the analytical data for the different physical-chemical variables. A series of lab-scale extractions, both with and without the addition of acid, was performed at standardized conditions ($T = 70$ °C , 3 grams of peel and, for the acidic extractions, 150 mL of 49 mM nitric acid). For each extraction, a total of two samples are taken at $t_1 = 20$ min and $t_2 = 240$ min, respectively. Samples were centrifuged at 10000 g for 10 minutes, and the supernatant diluted 1:10 with 0.3 M lithium acetate ($pH = 4.6$). The filtrate dilutions were analysed for pectin concentration (C^{bulk}_{pectin}) and intrinsic viscosity (IV) in a flow injection polymer analysis (FIPA) system. This system comprises a size exclusion chromatography column (150×7.8 mm, Thermo Fisher Scientific, MA, USA) with a triple detector array (TDA 305, Vicotek Corp., Houston, USA). A 0.3M lithium acetate ($pH = 4.6$) solution is used as eluent with a flow rate of 1 mL/min at 37 °C. The degree of esterification ($\%DE$) was determined by the 1H NMR method by Winning et al. (2007).

Near-infrared spectroscopy

Fourier transformed (FT) near-infrared (NIR) data were collected, on the citrus peel samples, using an ABB Bomem MB3600 FT-NIR technology spectrophotometer (ABB Bomem, Quebec, QC, Canada). The instrument was equipped with a rotating sample module with a quartz window. Spectral data for each sample were collected as the average of 62 single beam spectra at room temperature. The spectra were referenced against a white background spectrum (average of 62 scans). Samples were scanned over a 1000-2632 nm (resolution 16 cm^{-1}) range with a time interval of 20 seconds of measurement. Due to the presence of unwanted and unnecessary physical phenomena captured by the spectra it is often necessary to use preprocessing techniques in spectral data. The techniques provide mainly light scattering and baseline corrections. Spectral derivative techniques also help weighting more importance to otherwise unidentifiable variations within the spectra (e.g. small concentrations of analytes). Mean centering is always used after a sequence of preprocessing techniques (Rinnan et al. 2009).

Additionally, adequate variable selection (wavelength/wavenumber selection) is performed in order to achieve classification and predictive models with better performance. This selection prevents the capture of undesired variance between samples, as well as less redundant information, which will allow for less complex models (fewer latent variables needed to capture the necessary variance). Many different methods exist to assess the best variable selection (Xiaobo et al. 2010). One popular technique, which is used in this work, is called interval partial least squares (iPLS). This technique developed by Norgaard et al. (2000), builds separate local models on a number of (non-overlapping and equal width) subintervals of the full spectrum region. The regions are then selected based on the prediction performance of these local (and full-spectrum) models, by means of comparison of the root mean squared error of cross-validation.

4.1.3 Chemometric modelling

The data acquired in the food industry are quite complex and of different nature, often acquired from instrumental measurements comprising thousands of variables (e.g., each wavelength of spectra) for each sample. The complexity of the data has led to the need of diversifying the exploratory methods and employ multivariate data analysis to understand the information in the data. Chemometric methods can be put as a toolset of statistical techniques to analyse datasets with more than one variable (or type of variable). These variables are used simultaneously to perform exploratory (unsupervised learning), regression, or classification (supervised learning) analyses. All the methods considered in this paper have been exhaustively described and published elsewhere (e.g., Massart et al. (1988) and Sun (2009)); therefore, this subsection contains a summarized description of the chemometric methods employed to clarify and provide the reader with theoretical background and appropriate references.

Principal component analysis (PCA)

PCA is a method for dimensionality reduction of large multivariate data sets (i.e., spectral information), useful in many applications in the bioprocessing industries (Skov et al. 2014). In essence, PCA is a bilinear decomposition technique that, for a set of observations, relies on an orthogonal transformation of possibly correlated variables into a set of linearly (and mutually) uncorrelated variables called principal components. The transformation is defined such that the first principal component accounts for as much variance in the data as possible. The succeeding components will have the next largest variance, under the constraint that it is orthogonal to the preceding components. It summarizes the observation's information in fewer new variables. These new variables, named principal components, are thus composed of linear combinations of the original variables and constitute a set smaller than the original variables set. An original data matrix (X), is decomposed into a score matrix (T) and a loading matrix (P), with the residuals collected in a matrix (E): $X = TP^T + E$. The loadings define the new coordinates system, the weights that the previous/original variables have on each principal component (Li Vigni et al. 2013). The scores are the "amount of" those new artificial variables represented in particular sample, in other words, they are the coordinates of the samples in the principal component space (Bro and Smilde 2014).

Partial least squares regression (PLS-R)

PLS regression has been extensively used for multivariate regression modelling, especially applied to rapid spectroscopic measurements calibration with slow physical-chemical data (Rinnan et al. 2009). In a spectroscopic PLS-R application the purpose is to build a linear model between the desired response variable (y) and the spectrum (x), while concurrently maximizing the covariance between them by simultaneously decomposing the predictor and the response matrices iteratively into a reduced set of uncorrelated latent variables (LVs), thereby eliminating redundancy in the datasets (Geladi and Kowalski 1986). All models have been externally validated, by splitting the dataset into calibration and validation sets using the Kennard-Stone procedure (66% calibration; 34% prediction), keeping the NIR sample replicates together (Kennard and Stone 1969).

Partial least squares discriminant analysis (PLS-DA)

PLS-DA is a supervised classification method, where there is a requirement of prior knowledge, (i.e., the categories of samples). For this model identification, similarly to the PLS-R, a training set of samples for which the categories are known is necessary. This classification method is linear and based on the PLS algorithm, however, modified to perform classification. The main difference is related to the dependent variables (y) as these in PLS-DA are qualitative variables (Brereton and Lloyd 2014). In PLS-DA, a Y variable matrix is defined as a "dummy variable" matrix and it has

as many columns as the number of classes. The information about the class of each sample is provided through a binary code: all entries of each row (corresponding to a sample) are set equal to 0, except for the column corresponding to the category the sample belongs to, whose element is set equal to 1. If there are more than two classes, PLS-DA uses the algorithm PLS2. For each sample a prediction vector, with the size of the number of classes, is retrieved with values close to 0 and 1. The class is determined either by the maximum value in the y vector or by appropriate threshold setting for each class. Therefore, as in PLS regression, y dependent variables are predicted and thresholds (i.e., for a 2-class problem $y > 0.5$) can be defined to assign the sample to a corresponding class (Ballabio and Todeschini 2009; Bevilacqua et al. 2013). In PLS-DA the focus is not on the prediction error of the model but more the percentage of misclassifications obtained.

Cluster Analysis

Methods in this group of analysis have a common goal of finding groups/classes within a dataset, in which its members share more similarity to each other than with the rest of the observations, not in that group. An extensively used type of cluster analysis is hierarchical cluster analysis (HCA), which is based on distance measurement connectivity between observations. HCA is an unsupervised learning method, which can be agglomerative or divisive. An agglomerative method begins with each sample as its cluster and progresses agglomerating existing clusters into larger ones. Divisive methods start with a single big cluster containing all observations and are continuously separated progress by dividing existing clusters into smaller ones. All these methods require a distance measure between observations, with two popular ones being the Euclidean distance and the Mahalanobis distance. The latter is appropriate to account for multivariate directions. The different algorithms differ in the way the distance between existing clusters (inter-cluster distance) is defined and the decision guide for joining clusters (linkage rule)(Li Vigni et al. 2013).

4.1.4 Software

All chemometric calculations were performed using MATLAB ver. R2015B (Mathworks, Inc.) installed with the PLS Toolbox ver. 8.1. (Wise & Gallagher; Eigenvector Technologies).

4.2 Historical raw material variation assessment

Historically, the industry has dealt with raw material quality fluctuation in a heuristic manner, relying on the vast experience of their manufacturing teams and feedback information from production. Dialogue with raw material suppliers and process engineers has been essential in this way of manufacturing. It requires many years of manufacturing and supplier-relationship for the process engineers to have a “finger

on the pulse" and an intangible knowledge of what quality to expect from a certain supplier. This approach gives an "idea" of how to decide the process settings, but it is not fail-proof. Based on how certain categories (i.e., type of fruit, supplier) performed historically, the manufacturer is inclined to select a particular supplier for a particular pectin grade. For instance, the perception that lime peels provide the highest quantity of pectin and the higher end-product IV or that lemon peels typically have a higher end-product %DE, may provide the manufacturer with a false sense of security when purchasing raw material (May 1990).

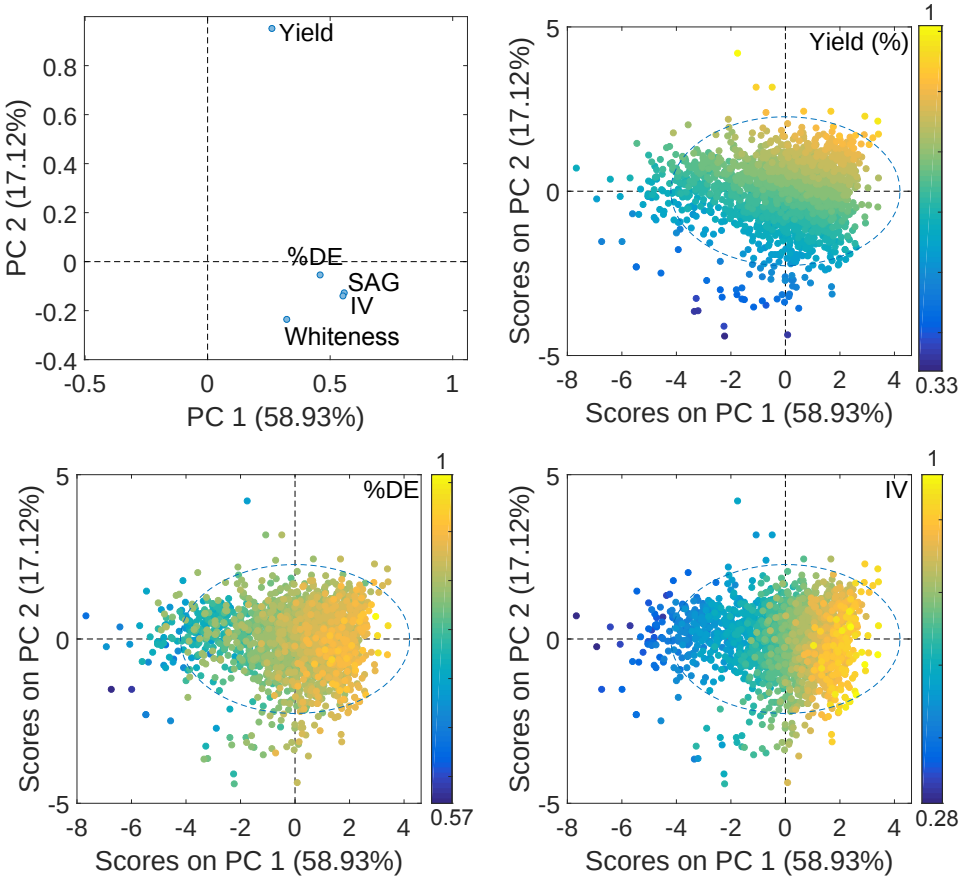

Figure 4.1: (Top Left) PC1 versus PC2 loadings plot. (Rest) PC1 versus PC2 score plot, colored by the indicated variables. All variables were standardized to their highest value in the dataset.

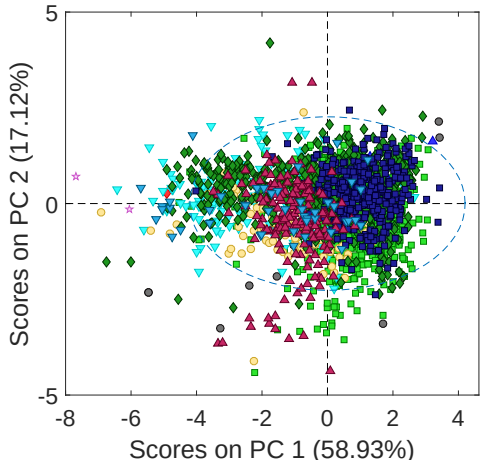

Figure 4.2: PC1 versus PC2 score plot, colored by country.

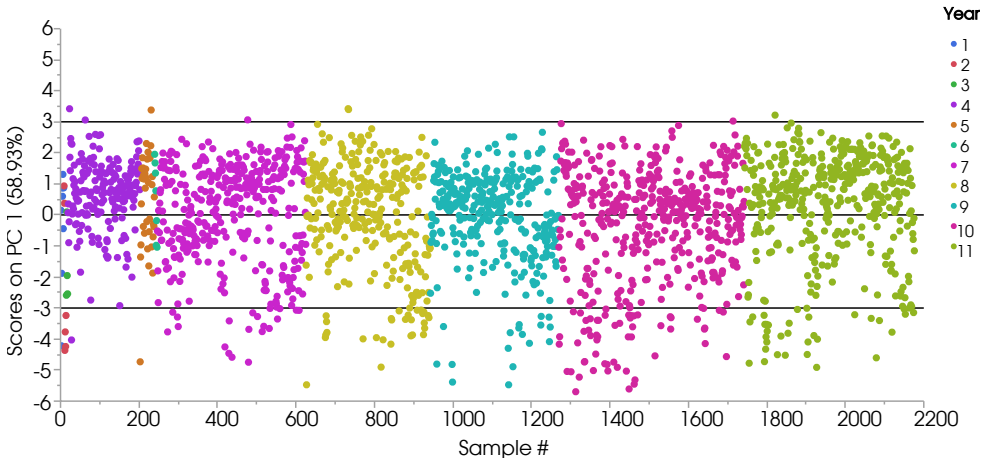

Figure 4.3: Samples PC1 variation throughout the years in the dataset analysed

To assess just how influential the peels are in terms of end-quality of pectin, an exploratory multivariate analysis by means of principal component analysis is made on the dataset described in Section 4.1.1. A representation of the two first principal components can be seen in Figure 4.1. According to the loadings plot and PC1 mostly separates the samples with respect to the functional variables, while Yield(%) contributes mostly to the variance captured in PC2. Samples with high values for Yield are situated in the 1st quadrant and decrease diagonally across the plane where the lowest values are situated in the 3rd quadrant. IV and %DE are almost uniquely

separated by PC1. Samples with high values for these variables will be found in the 1st and 4th quadrant. Ultimately it can be said that samples situated in the 1st quadrant of the PC1xPC2 plane encompass the peels which have the most desirable attributes. Significant ranges are observed for the variables yield (0.33-1), %DE (0.57-1) and IV (0.28-1).

Citrus peels are purchased depending on the availability and price from various countries and suppliers, with the same suppliers displaying significant fluctuation in manufacture performance for different cultivar years. This is assessed in Figure 4.2, in which the same principal component scores are colored by country of origin. Even though a few countries have prevalent presence in the PC1 positive range, it is still visible that the spread in the plane for samples of the same country is appreciable. This represents a wide range deviations in the yield, %DE and IV of end-product pectin from suppliers from similar background. In Figure 4.3 it can be seen that over the course of eleven years, the fluctuation in IV and %DE is something which should be expected.

An approach to operate a plant based on this feedback information with the lack of systematic critical material attributes determination, requires a long period of time to acquire the experience necessary to operate satisfactorily. Once the expertise exists, the guidelines on how to operate for a specific peel can be readily applied, which means a lower in-operation time effort. However, in this approach the critical material attributes are still unmeasured and thus uncontrolled noise factor in the process. This analysis shows that variations due to the use of different peels in the process are inevitable. Without an approach to estimate critical material attributes information about the raw material, the producer will not be able to cope accordingly in the face of eventualities (e.g., supplier bankruptcy) that require a sudden change of the main supplier, or even if the raw material quality of a trusted partner changes due to changes in their protocols.

4.3 Critical material attribute assessment approaches

There are no specific regulations for the assessment of raw materials, but there are guidelines in the context of good manufacturing practices of active pharmaceutical ingredients. The guidelines push for the evaluation of quality and establishing of acceptance criteria in raw material, correct labeling and documentation of end-product deviations to detect changes resulting from modifications in raw materials (ICH Expert Working Group 2000). An overview of five different possible approaches for quality assessment of peels is illustrated in Figure 4.4, of which the last three (highlighted with a dashed line) will be assessed comparatively in a quantitative analysis on their uncertainty. The scheme shows the different approaches regarding the statistical information that can be gathered as well as a qualitative time-effort relative comparison regarding development, as in how much time it is necessary to allocate before being able to use the approach, and in-operation use, as in how time expensive is the approach during production. Figure 4.4 has elements specific for the pectin case,

but transversal for other bio-raw material cases where key parameters of the raw material are identifiable and PAT is applicable. This section will discuss each approach in the context of the article's case study, addressing the necessary developments for each approach:

- Descriptive statistics of reference variables:$u_{peel} = [C^0_{pectin}, C^0_{protopectin}, IV_0, \%DE_0]$
- Exploratory Analysis and Data-visualization of both spectral and reference data
- Cluster Analysis
- PLS-R and PLS-DA modelling development for u_{peel} variables

4.3.1 Wet-Lab Analysis

The following approach in Figure 4.4 requires the determination of critical material attributes measurable via experimental analysis. The development of an analytical method starts with the definition of the desired characteristics to assess in a substance. Once identified the relevant critical material attributes, the analytical procedure depends on the choice of sample preparation, analytical instrumentation and methods that are appropriate for the nature of the sample and the intended goal of the analysis (Chauchan et al. 2015). After development, it is necessary to evaluate

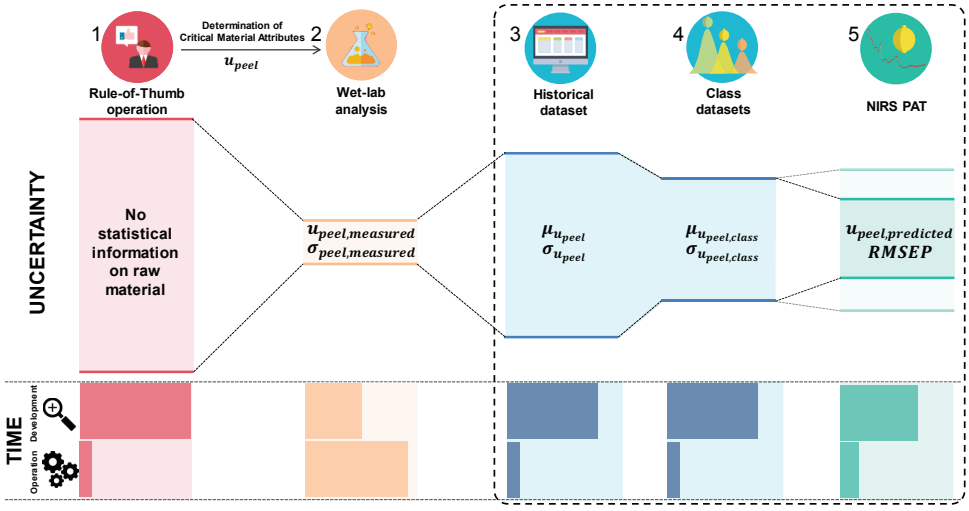

Figure 4.4: Conceptual comparison of raw material characterization approaches qualitatively regarding uncertainty information gathered (and expected relative span) and time-effort it would require the manufacturer to implement the approach and its in-operation time expenditure. The box highlights the three approaches under quantitative comparison in this study.

the method by setting appropriate acceptance criteria in validation experiments for typical parameters, such as specificity, linearity, accuracy, precision, range, detection limit, quantitation limit, and robustness (Green 1996). Although bio-analytics are fully regulated in the pharmaceutical industry (ICH Expert Working Group 2005; G. Tiwari and R. Tiwari 2010), the food and bio-based fields lack detailed guidelines for analytical method validation.

The process of identifying the desired attributes, development of the analytical method and validation can be long and require a significant amount of experimentation to achieve robustness. However, one can expect to achieve an acceptable method in less time than it takes to identify consistently (if possible) the quality distribution in the supplier market by qualitative feedback from production (approach in the *Rule-of-thumb operation* section). While in operation, these methods can be extremely time and resource consuming (e.g., personnel, reagents), with the results, typically, only accessible after hours (even days), preventing the use of the information obtained for control of the current process. The estimate resulting from a bio-analytical method is expected to have a relative standard deviation of less than 15% (ICH Expert Working Group 2005).

Determining raw material attributes that are parameters (or initial states) of a model is highly useful as it enables their direct use. The outputs of the analysis would be inputs in the model, creating a flexible tool for simulation of key performance indicators for a given raw material and process settings. Different models for pectin solid-liquid extraction have been developed (Cho and Hwang 2000; Durán et al. 2015; Minkov et al. 1996; Pagán and Ibarz 1999), but lacked the integration of critical material attributes in the process dynamics. Andersen et al. (2017) have developed a model which describes the relevant process key performance indicators, pectin concentration in the bulk media (C^{bulk}_{pectin}) effect of temperature and pH and incorporates parameters which are both related to the concentration, C^0_{pectin} and $C^0_{protopectin}$, and initial quality, IV_0 and $\%DE_0$ of pectin material in the peels. These parameters, $u_{peel} = [C^0_{pectin}, C^0_{protopectin}, IV_0, \%DE_0]$, constitute a vector of critical material attributes which will vary from peel to peel. The experimental procedure and analytical methods described in the *Chemical reference* section were developed as a means to measure these parameters directly:

- $C^0_{pectin} = C^{bulk}_{pectin}(t_2, \text{water extraction})$, where t_2 is the second sampling time and, assuming only readily available pectin is extracted with water.
- $\%Yield = \frac{C^{bulk}_{pectin}(t_2, \text{acid extraction}) \cdot V_{extraction}}{m_{peel}} \times 100$, which is not explicitly a u_{peel} parameter, but it is used as a variable in the results for the ease of dimensionless comparison between samples.

indirectly:

- IV_0 and $\%DE_0$ by assuming a first-order reaction (i.e., $ln[IV] = ln[IV_0] - k \times t$ and calculating the y-intercept with the IV and $\%DE$ data at t_1 and t_2.

moreover, by reconstruction:

- $C^0_{protopectin} = C^{bulk}_{pectin}(t_2, \text{acid extraction}) - C^0_{pectin}$, reconstructed from the measured variables above and thus not used in the results below.

Performing these assays will provide us the uncertainties associated with the measurements for each peel. Lab-scale extraction methods are favored in detriment of the pilot extractions. It is (manifold) less resource and time intensive, and the extraction conditions can be carefully standardized at lab scale (e.g., thermal bath, perfect mixing, and better sample handling). However, these experiments are still time and resource consuming methods. This is undesirable for a standard operating procedure, aimed at systematic use in production. There is a need to streamline the assessment of peel quality, but maintaining the statistical information essential for further applications.

4.3.2 Dataset statistics

Once the critical material attributes are determined, the method is established, and data has been continuously collected it is possible to characterize the raw materials statistically. If the information is gathered on a representative population, including prospective raw materials not currently in production, the approach offers a robust variation assessment of the tested materials. The crucial point in this approach is the need to guarantee that the samples tested do cover the expected variation in the production environment. When established, information on the critical material attributes of incoming raw material can be inferred from the up-to-date dataset statistics, and routine sample analysis can be made to update the dataset. Building this dataset, as the third column in Figure 4.4 illustrates, solves the in-operation time pitfall of the previous approach. However, the process to ensure representativity can be lengthy and resource consuming there is no guarantee that the extremes are investigated, and future samples are out of the analysed limits (Skibsted 2011).

Table 4.1 offers the statistical summary of the measured variables in-peel. Due to the happenstance nature of the dataset, not all samples have a complete critical material attributes vector (u_{peel}). Each variable in Table 4.1 has indicated the number of samples, n, which got measured for that particular variable. The summarization of the data is assessed using empirical distributions: a measure of the central tendency (mean, median, mode); a measure of spread (range, quartiles, standard deviation); a measure of asymmetry (skewness) and peakedness (kurtosis) of the data distribution. Skewness measures the lack of symmetry in distribution. A variable is symmetric if it looks the same to the left and right of the center point. Kurtosis measures how tailed a distribution is relative to a normal distribution. A normal distribution is symmetric with well-behaved tails. For this type of distribution, the skewness is close to 0 and kurtosis has the value of 3 (Tukey 1977). Distributions with higher kurtosis will have heavier tails (possible outliers). The variables in Table 4.1 display similar behavior to a normal distribution. However, operating based on the complete

Table 4.1: Empirical measures of central tendency, spread, and asymmetry of the measured critical material variables. Full dataset included (n is the number o samples for each variable, μ the mean, σ the standard deviation and p10-p90 are the percentiles).

	$\%Yield$	$\%DE_0$	IV_0 (dL/g)	C^0_{pectin} (g/L)
n	68	63	68	80
μ	24.98	73.50	7.78	1.59
σ	2.04	2.81	1.14	0.59
min	19.81	67.47	5.44	0.34
max	30.97	78.27	10.10	3.01
p10	22.12	68.85	5.89	0.77
p25	23.69	72.18	7.24	1.23
p50	25.14	73.83	7.97	1.61
p75	26.22	75.53	8.60	1.95
p90	27.67	77.03	9.36	2.43
skewness	0.13	-0.52	-0.28	0.27
kurtosis	3.28	2.6	2.44	2.79

dataset statistics could be flawed and does not provide the flexibility to account for significant raw quality variations which may occur. The uncertainty associated with the raw material might be too large to consider a single statistical distribution for valuable implementations.

In the citrus peels example, this is known a priori because we know beforehand that different fruits are typically used to produce pectins with different specifications (May 1997). A mean of 7.78 and standard deviation of 1.14 for IV_0 places the manufacturer's initial guess for a given peel within the operational range of the two different product-specifications previously mentioned: jam ($5 \leq IV \leq 6$ dL/g) and jelly ($IV \geq 6$ dL/g). However, many peels in the dataset have a measured $IV_0 < 6$ dL/g and the negative skewness (-0.28) indicates that the distribution for IV_0 has higher incidences for values $IV_0 < 7.76$ dL/g. This would be problematic if we adopted the global mean of the dataset value and attempted to produce jelly-type pectin.

4.3.2.1 Class dataset statistics

The shortcomings in the previous section motivate the search for classification that yields a narrower window of uncertainty regarding the critical material attributes. Moreover, they can provide the manufacturer with specific operational guidelines based on the class of the incoming raw material, connecting different optimal operating setting to different raw materials. These classes can be defined with either quali-

fied expert-knowledge (discriminant) classification or through unsupervised learning and clustering algorithms.

Expert-knowledge classification

Expert-knowledge classification relies on prior information, which the manufacturer uses to sort different samples, for example, attributing classes to raw materials based on their supplier, country of origin and in this case the type of fruit. Figure 4.5 shows what this classification yields in statistical terms using box-plot visualization. This graphical method provides additional information to the summary statistics. It can identify outliers, changes in the data distribution across different groups and variables and even highlight relationships between variables.

As it can be seen in Figure 4.5, there are a few outliers in the selected groups and variables, the most noteworthy being the orange sample which has outlier values for IV_0 and $\%DE_0$. This is a case where it is possible that an error while compiling the dataset occurred in the labeling, as it is evident that this sample has a very distinct value in variables where its group (orange) is well differentiated from the rest. The sample could have been a lime or lemon, with a higher probability as the former since it most closely fits the lime interquartile range (the box in Figure 4.5 that represents 50% middle) in all variables.

The box-plots highlight the dissimilarities between fruits, with some variables having more discernible differences between fruits than others do. For instance, the IV_0 and $\%DE_0$ box-plots clearly isolate orange and the rest (see Figure 4.5). Lemon and lime, albeit overlapping in their data range, still have different central differences. The lower IV_0 from orange supports the study by Kaya et al. (2014) that claims oranges contain longer or more numerous side stretches which provides a flexible conformation, leading to decreased intrinsic viscosity values compared to lime and lemon. For these quality variables, the fruits reveal an almost normal distribution (mean and median close together, and the whiskers have similar lengths). This is not the case for initial pectin content in the peels.

For the C^0_{pectin} variable, data distribution by fruits overlaps significantly with no highly discernible central distribution difference between fruits. However, Figure 4.6 shows visible clustering by the supplier. This indicates that the peel pre-treatment of each supplier is a defining factor, rather than the fruit group itself. This is in agreement with what is stated in the literature, that the pectin solubility is promoted by a combination of the de-esterification of the polygalacturonic acid backbone (pectin methylesterases) facilitating the depolymerization of pectins (polygalacturonase) and the cleavage of linkages between side chains of pectin and hemicelluloses. This conversion of protopectin to soluble pectin is dependent on the pre-treatment (blanching, washing, drying, etc.) applied by the supplier (Lopez da Silva and Rao 2006).

The different stages of fruit maturity, when the peels are collected and sold by the supplier, also contribute to the differentiation between pectin in the different peels (Sriamornsak 2003). Adopting a classification based on the fruit type is a lesser than an optimal solution since for the u_{peel} variables there is excessive overlap between the

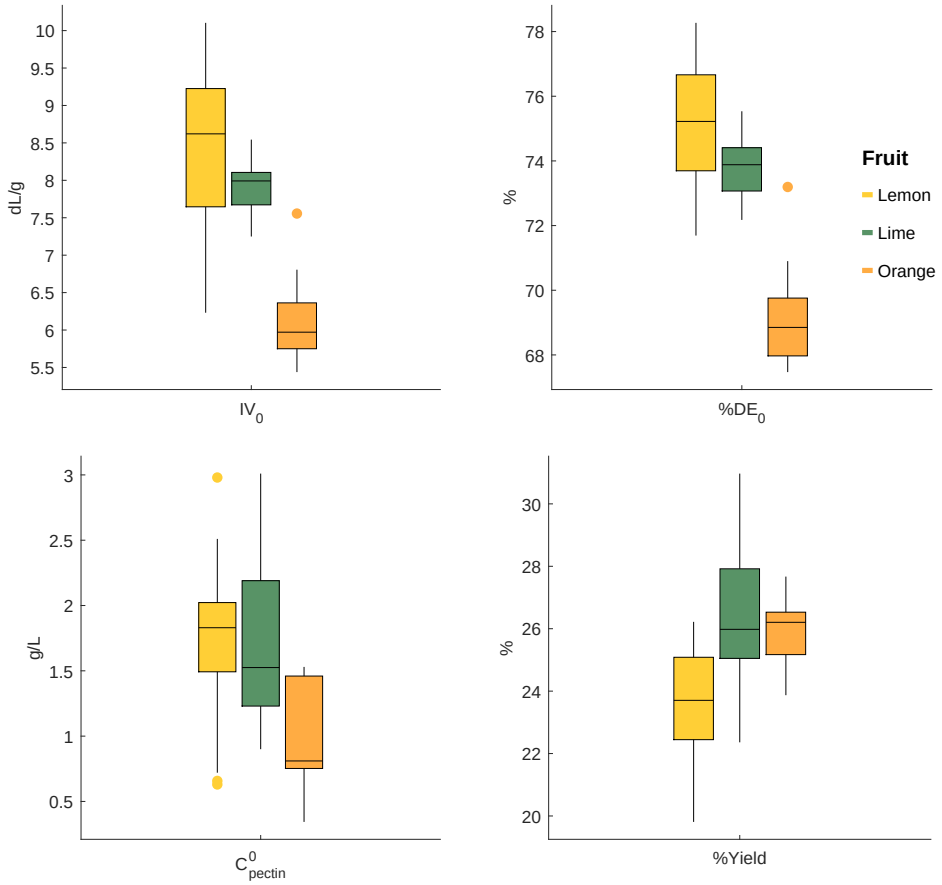

Figure 4.5: Boxplots of samples grouped by lemon, lime and, orange for the measured response variables $\%Yield$, $\%DE_0$, IV_0 and, C^0_{pectin}.

fruit types. This motivates the examination for classes that are "blind" to the fruit type.

Unsupervised learning and cluster analysis

Unsupervised clustering is performed with little or no information about class structure before the classification; the classes form based on the distance of the u_{peel} vector between samples. A PCA model, based on the measured critical material variables, with two components (84% cumulative variance), is built and analysed such that the loadings and scores can be visualized and can be interpreted together in a biplot (Gower 1995). The PC1 vs. PC2 bi-plot is seen in Figure 4.7. The first principal

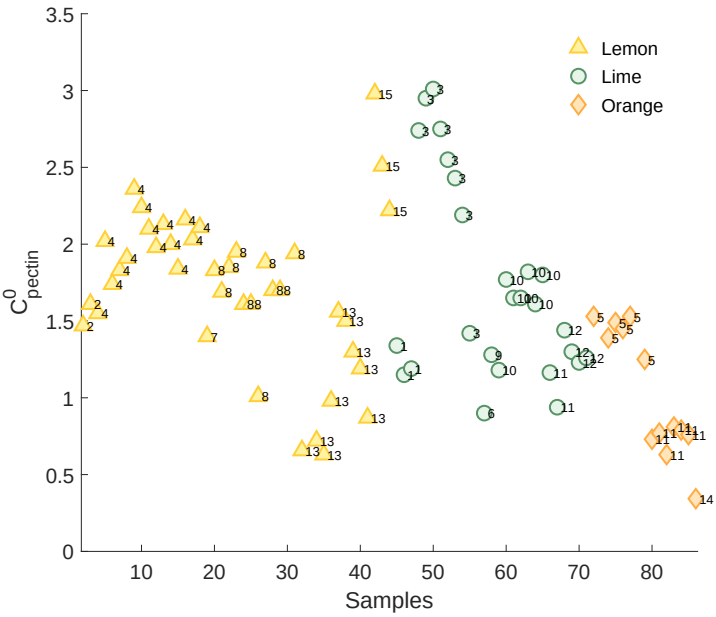

Figure 4.6: C^0_{pectin} scatter plot. Samples colored by lemon, lime and, orange. Number labeled by supplier.

component is able to make a separation between the three fruits. This confirms what was previously assessed in the univariate analysis, that the fruits were distinguishable to a certain degree with the u_{peel} variables. The fruits are mainly separated by the IV_0 and $\%DE_0$ variables. However, there are lemon and lime samples that the PCA model cannot tell apart. This is in accordance with the previously assessed information on the overlaps between lime and lemon. Other information extracted from Figure 4.7 is that the IV_0 and $\%DE_0$ variables are near each other and far from the origin, meaning they are correlated (with respect to the variation explained by the components). Cluster Analysis of these scores will allow for establishing classes that respect the closeness of data in a multivariate sense, irrespectively of their fruit type.

In Figure 4.8A, the classes originated from using Ward's agglomerative algorithm method(Murtagh and Legendre 2014), using Mahalanobis distance, on the PC1 and PC2 scores from the previous PCA model. This approach generates more homogeneous classes, which are not overlapping in the PC1 vs PC2 plane, contrary to the fruit classification. Class 1 has both lime and lemon samples as their constituents. Class 2 is comprised of only lemons, which differentiate themselves from the rest for their high IV_0 and $\%DE_0$ values. Finally, class 3 coincides entirely with the orange samples, which are known from the univariate analysis to be distinguishable from the

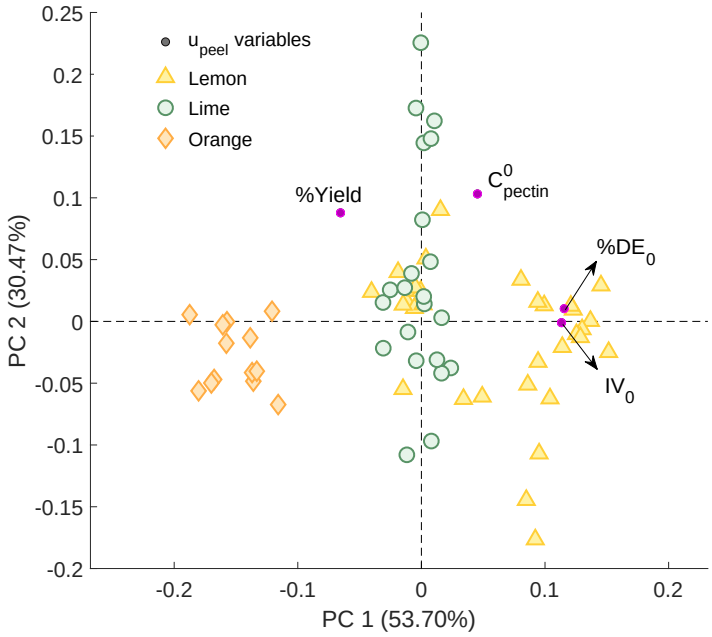

Figure 4.7: PC1 vs. PC2 biplot. Colored by lemon, lime and, orange, with the u_{peel} variables labelled.

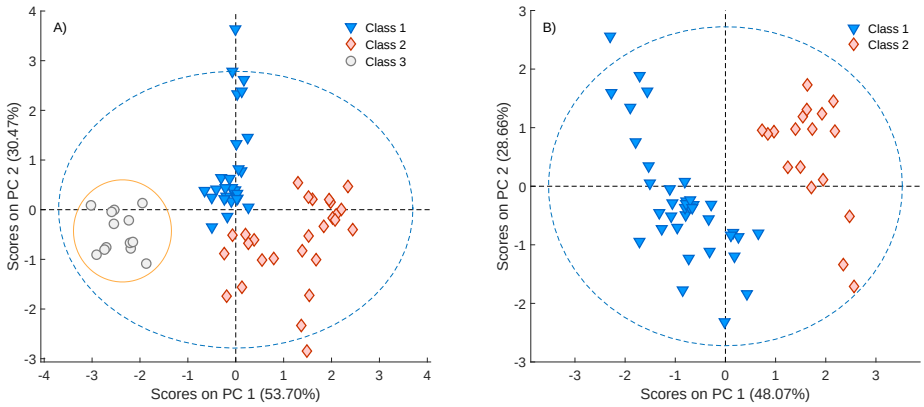

Figure 4.8: PC1 vs. PC2 score plots. Colored by cluster classes obtained with Ward's agglomerative algorithm (Mahalanobis distance). A) Includes all 69 samples B) Excludes the grey samples highlighted in A). The cluster analysis is repeated after the exclusion of the grey samples.

other fruits, particularly in terms of IV_0 and $\%DE_0$ values. It is a spurious classification attempt since the experimenter can identify if a sample belongs in class 3 merely by identifying the sample as an orange. This encourages the clustering in Figure 4.8B within the lemon and lime samples, in an identical mode to the previous clustering but without the undesired variance that is captured by adding the orange samples. The classes in the new analysis separate once more the samples which have high IV_0 and $\%DE_0$ values (class 2) from the others. This yields class 1, with mixed samples of lemon and lime, while class 2 is exclusively composed of lemons. To determine the class of a sample the determination of its u_{peel} vector (or partly) is required, which can still be cumbersome for the operational decision-making and optimization in a timely manner. Any monitoring or process optimization strategy will opt for an approach that provides the best estimates (with the least uncertainty) in a rapid manner. This encourages coupling the critical material attributes, u_{peel}, with a spectroscopic method.

4.3.3 Spectroscopic coupling

Spectroscopic techniques provide the manufacturer with a fast and, in most circumstances, non-invasive and non-destructive tool. The use of spectroscopic techniques can be fruitful in reducing drastically the in-operation time when compared to wet-lab analysis, and providing the manufacturer with a better estimate of the critical material attribute than the dataset statistics. This is illustrated in the last column in Figure 4.4. These tools can also reduce the time-effort put into developing the quality assessment approach if used for the initial screening of the different batches of raw material. The analytical tests can then be made in a reduced set of the original selection, carefully selecting the materials that cover the largest variation ensuring a representative dataset. In the following examples, the methods are used in the context of the full-dataset employed in the previous approaches. Within the possible techniques, NIRS is a notably reproducible and robust spectroscopic method that has proved its rapid non-invasive use across the food and agrochemical industries. Previous studies register NIRS capabilities for detecting pectins and pectin quality parameters.(Baum et al. 2017; Engelsen et al. 1998)

Principal component analysis is performed to investigate the samples separation based on their full NIR spectra information. The score plot for the two first components is shown in Figure 4.9, with the spectra being pre-processed with the common standard normal variate (SNV) and mean centering techniques. Pre-processing attempts to remove physical variability (scatter correction), so that the samples can span in the principal components space due to variations in the chemical matrix. The first component explains 83.7% of the variation in the pre-processed NIR samples. By analysing the PCA score plot in Figure 4.9A, it is possible to observe a discernible gap between samples in PC1 (highlighted with a red box). The reason for this can be assessed by evaluating the spectra, in Figure 4.9B, together with the loadings plot for PC1, in Figure 4.9C. A wavenumber range which has high impact in the separation

across PC1 is 8400-10000 cm^{-1}, in the third overtone NIR region. Not only it is the region with the weaker intensity it can also be seen that it manifests specular reflectance effects with no sharp structure. The first component is largely composed of effects resulting from differences in sample pre-treatment (grinding, storage, etc.) rather than differences in the pectin molecules and its availability in the peel. This large variation captured only adds noise to the purpose of using the spectra to infer u_{peel} information from a sample's spectrum. This is an indication that this region should not be used for further classifications or predictions of the physicochemical variables of pectin in the peel matrix. In succession, the second component has dominant loadings in a band that corresponds to RCO2R' ester-groups (maximum at 5161 cm^{-1}). This is a good indication that the samples are separated by their degree of esterification in this direction. It is important to note that the replicates are also close to each other, indicating a good degree of robustness. Another interesting feature is that the physical effect is not exclusive for a single fruit (1 lemon sample) and only the suppliers "1", "3", "10" and "13" are affected, but rather than being a characteristic spectroscopic fingerprint from the supplier, it is possible that it is merely the effect of all these samples having been pre-treated similarly. One evidence for this possibility is that it is possible to find other samples of supplier "13" outside the highlighted box in Figure 4.9A. In these components, samples from the same suppliers do not necessarily occupy the same regions of the PC space. For example, this is also visible for the supplier "5".

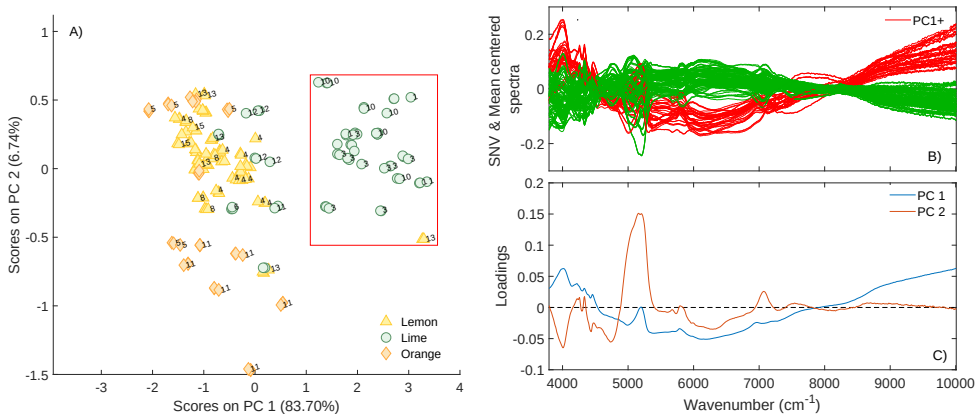

Figure 4.9: A) PC1 vs. PC2 score plot colored by lemon, lime and, orange, number labeled by the supplier. B) SNV and mean centered spectra. PC1+ indicates the samples highlighted in the box in A). C) Loading plots for the first two principal components.

Classification aiding tool

PAT based classification can overcome the analytical effort pitfall, through spectroscopic class assignment or spectroscopic prediction of classes assigned based on reference analysis measurements. Based on the unsupervised classes obtained from the cluster analysis in the *Unsupervised learning and cluster analysis* section, a PLS-DA model attempts to predict these classes (obtained with the reference data proximity) only by using the NIR spectra of the samples. By predicting the class to which a sample belongs to, the statistics (mean, standard deviation) from that class can then be used as the u_{peel} vector for that sample, the same way the statistics for the supervised classes in "Fruits" would be applied. The spectra are preprocessed with the same techniques applied in the analysis above, standard normal variate (SNV) and mean centering.

Cross-validation is applied by leaving out the NIR replicates of the same peel together, thus avoiding over-fitting of the data. A confusion matrix is used to assess the PLS-DA classification success in classifying the samples, illustrating the correctly classified samples, type-I errors (false-positives), and type-II errors (false-negatives) samples for each category in the matrices presented in Table 4.2. It can be observed that 16 (Class 1) and 18 (Class 2) samples were misclassified by the model attempting to classify all samples. It can correctly identify class 3 samples (oranges), but there are errors on samples close to the intersection between the two classes containing both lemons, and limes. This indicates that the model is not perfect (fail proof efficient) in the classification process. However, a practical implementation for our case, making use of a priori information of the fruit type, would be to classify the lemons samples between the two classes in Figure 4.8B. This reduces the task complexity of the model dramatically and yields an almost perfect cross-validated result, seen in Table 4.2. The two misclassifications are replicates of the same peel, with a third replicate having been correctly classified. Furthermore, by performing a variable selection with an iPLS algorithm the cross-validated result is improved for both classification models, with the latter achieving a perfect cross-validated result.

Critical material attributes prediction models

If the raw material key parameters are liable to be calibrated with a spectroscopic tool, an at-line PAT application can provide that timely information. The calibration makes use of the acquired dataset in the attempt, of quantitatively predicting the u_{peel} variables. The same calibration and external validation sets were used for all models. The results from the $\%DE_0$ model, illustrated in Figure 4.10, showcase the adequacy of the NIRS technique for the characterization of the specified material attributes. This model was built in an iterative fashion, trying different combinations of spectra pre-processing and variable selection. The outcomes from the different models are compared based on their performance to predict the same external validation set. This process follows these steps:

Table 4.2: PLS-DA classification confusion matrix for a cross-validated set (keeping the spectral replicates together) for both scenarios of classes presented in Figure 5. Both models are built with full spectra and 10 latent variables.

		Actual Classes		
		Class 1	Class 2	Class 3
Predicted	Figure 5A (All Samples)			
	Class 1	61	18	0
	Class 2	16	56	0
	Class 3	0	0	38
	Figure 5B (Lemon samples)			
	Class 1	37	0	-
	Class 2	2	50	-

1. Choosing the pre-processing of spectra (i.e., SNV + 1^{st} derivative + mean centering) and lab-reference $\%DE_0$ samples (i.e., autoscaling)
2. Variable selection (i.e., no variable selection: full spectra)
3. Number of latent variables selection. Register the model performance.
4. Re-do model with subtle changes to variable selection (i.e., use iPLS algorithm). Register the model performance.
5. Go through sequence 1-4 with changes to 1) (i.e., SNV + 2^{nd} derivative + Mean Centering), register the model performance.
6. Comparison of root mean square error of prediction (RMSEP) of the external validation set

From the different modelling iterations, the $\%DE_0$ model that yielded the best results was obtained for the regression with a pre-processing of the spectral data with SNV + 1st derivative + mean centering. The R^2 is the squared correlation coefficient providing the explained y-variance and $bias = (y_{pred} - y_{ref})/n$, with n as the number of samples. There is a high correlation in the model and the normalized-RMSEP ($nRMSEP = \frac{RMSEP}{(X_{\text{obs,max}} - X_{\text{obs,min}})} \times 100$) is close to 10%. The other models and its specific use of variables are summarized in Table 4.3. There is, however, a relatively high bias, which results from a poorer predictive capability of the model for orange peels. It should be pointed out that at this point, the distribution of the samples (see *Citrus peel samples* section) considered for this model building is somewhat skewed and might not allow for a generic model that predicts critical material attributes for all peels adequately. Further efforts will go into validating and consolidating the

results. Nonetheless, the study showed in principle that the technique is capable of prediction.

Table 4.3: PLSR models summary table. All models were pre-processed with SNV + Savitsky-Golay (1st derivative, 2nd order polynomial, 7 points) + mean centering

	Variable Selection	Latent Variables	RMSEP	R^2
$\%DE_0$	iPLS 3793.3-6613.4 cm^{-1}	6	1.05	0.9
IV_0 (dl/g)	Full-spectra 3793.3-10007.9 cm^{-1}	6	0.58	0.82
C^0_{pectin} (g/l)	Full-spectra 3793.3-10007.9 cm^{-1}	8	0.2	0.84
Yield (%)	Full-spectra 3793.3-10007.9 cm^{-1}	6	1.57	0.53

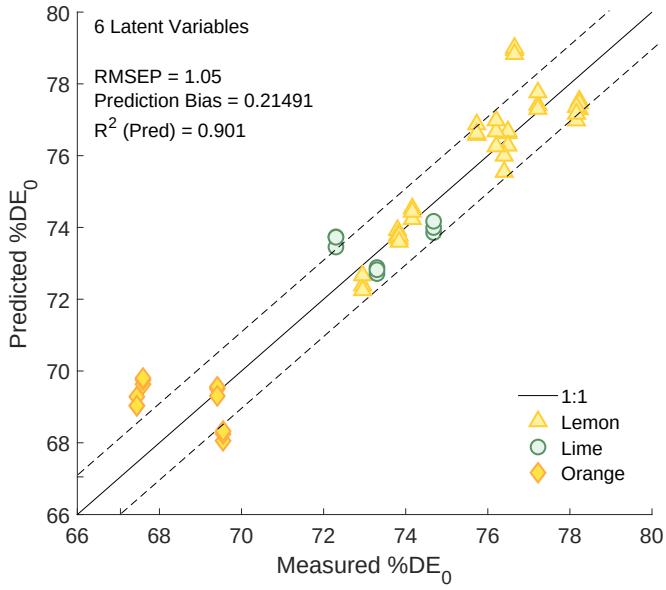

Figure 4.10: Predicted vs measured $\%DE_0$ plot. Colored by lemon, lime and, orange. The dashed line is equivalent to 1:1 fit ±RMSEP. The models for the other variables are not plotted but their performance is registered in Table 4.4

4.3.4 Discussion

A quantitative comparison of the three approaches discussed previously i.e., fruit classes (*Expert-knowledge classification* section), cluster classes excluding the orange samples (*Unsupervised learning and cluster analysis* section) and the partial least squares regressions (*Critical material attributes prediction models* section) is shown in Table 4.4 and Figure 4.11. The comparison is made by a measure of central tendency (i.e., the mean μ) and a measure of spread (i.e., the standard deviation σ) within the same dataset context. When taking the mean of a variable as the prediction for a certain class, the uncertainty on such assumption can be defined by the standard deviation of such variable in that class. For example, if the manufacturer operates based on fruit discrimination and a lemon peel arrives in production, they can assume the values in the vector $\hat{u}_{\textbf{incoming lemon peel}} = [8.4; 75.1; 1.72; 23.69]$ as the critical material attributes profile for that raw material. It is then possible to optimize the process stochastically according with the desired product specifications, knowing that the uncertainty associated to such assumption can be defined by the standard deviations of the $\hat{u}_{\textbf{incoming lemon peel}}$ variables: $\sigma_{\hat{u}_{\textbf{incoming lemon peel}}} = [0.98; 1.98; 0.51; 1.57]$. It can be assessed in Table 4.4 that there is a statistical uncertainty improvement from fruit classes (qualified expert-knowledge) to cluster classes (unsupervised). Across all u_{peel} variables, the cluster classes have a smaller maximum standard deviation, when compared to fruit classes. The uncertainties associated with the model predictions are well in the range (i.e., IV_0 and $\%Yield$), or smaller (i.e., $\%DE_0$ and C^0_{pectin}) than the minimum standard deviations obtained in the group statistics. In Figure 4.11, the mean value of the relative standard deviations (also known as coefficients of variation) of the composing classes of a given approach is calculated for each u_{peel} variable.

The quantitative comparison shows how both statistics of historical analytical data and the performance of predictive spectroscopic models are useful to assess raw material quality with an appropriate uncertainty. The manufacturer can opt to rely on the class statistics or implement the PAT tools, depending not only on performance but also on practicality and economic feasibility. For the fruit-based classification, experimentation can be skipped once a representative dataset is available. Samples are classified given an intrinsic characteristic which is known *a priori*. Classification of samples based on their original fruit provides a decent distinction and requires no classification model. However, it shows to be insufficient for specific critical material attributes, i.e., C^0_{pectin} (see Figure 4.6), which show an overlap of different fruits.

Alternatively, classification of samples based on cluster analysis joins samples in groups based on their multivariate proximity regarding the identified critical material attributes. This yields classes that are more homogeneous concerning uncertainty (see Figure 4.8 and Table 4.4). This approach makes the most of all available wet-lab features of the samples and is not influenced by biased classifications based on heuristics. This results in classes which include a mix of fruits. However, this type of classification requires additional information. Performing the full lab experimentation would allow for this classification, but is also beneath the purpose of bypassing the wet-

Table 4.4: Comparison of the three raw material quality assessment approaches

	Fruits			Cluster Classes			
	Lime	Lemon	Orange	1	2		PLS-R
IV_0							
μ	7.92	8.4	6.01	7.74	9.12	R^2	0.82
σ	0.31	0.98	0.41	0.48	0.47	RMSEP	0.58
$\%DE_0$							
μ	73.85	75.1	68.71	73.59	76.55	R^2	0.9
σ	1.00	1.98	1.15	1.04	1.11	RMSEP	1.05
C^0_{pectin}							
μ	1.72	1.72	1.11	1.81	1.69	R^2	0.84
σ	0.64	0.51	0.36	0.59	0.56	RMSEP	0.2
$\%Yield$							
μ	26.22	23.69	26.03	25.57	23.02	R^2	0.53
σ	2.07	1.57	0.96	2.00	1.37	RMSEP	1.57

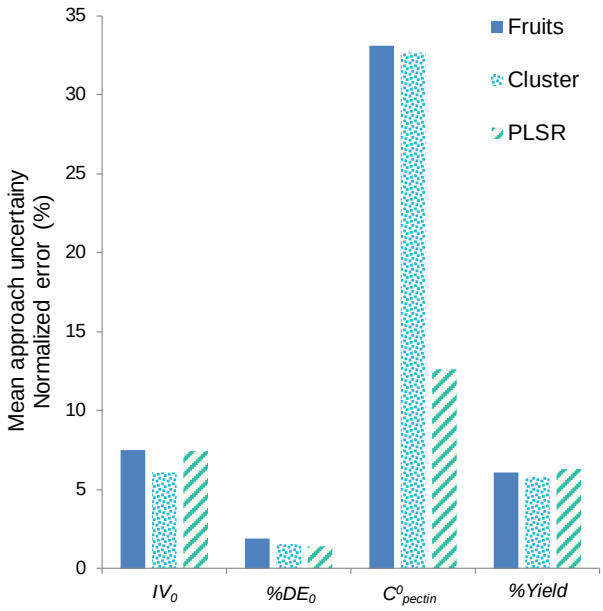

Figure 4.11: Average coefficient of variation (relative standard deviations) for the three approaches in comparison.

lab analysis. A way to partially circumvent this is to attempt classification through partial experimentation and posterior determination of the lacking variables through correlation or missing-data algorithms.

The *Spectroscopic coupling* section explores the use of NIRS as a tool for rapid identification of these classes. In a previous study by (Engelsen et al. 1998), NIRS showed discriminant capability and was able to successfully distinguish citrus peels samples originating from different countries and specific producer fingerprints. A particularly cumbersome difficulty is how the raw material sample pre-treatment affects the multivariate applications. This is observed in Figure 4.9, where a great part of the variance between samples captured is non-dependent on the chemical matrix of the sample. This is a common case in biological samples where scattering properties are complex (Rinnan et al. 2009). Caution is necessary for standardizing the pre-treatment of samples. Different sample grinding could also induce differences in the chemical matrix, specifically in the water content. Moisture loss occurs during grinding, mainly due to air throughput exposure in the particle which has a bigger superficial area. The grinding also exposes the particles to overheating. In this study, the approach with PLS-DA modelling suffers from a few misclassification errors between class 1 and 2 (Table 4.2). However, a faultless classification is possible when combining heuristic information (knowledge of a samples fruit) and the PLS-DA model on lemon samples. This application has a smaller classification task and is still very relevant in operation as it allows the users to identify via NIR if a lemon belongs to class 1 or class 2 (and expect higher IV_0 and $\%DE_0$ values). With the creation of this class distinction, the user avoids the underestimation, by assuming a fruit class, of IV_0 and $\%DE_0$ values for good performing lemon samples. It provides not only better precision, as seen in Table 4.4 and Figure 4.11, but also better accuracy. An alternative approach would involve building classes based on proximity (clustering) of the near-infrared samples PCA scores. This would enable better performance of a classification model based on the NIRS. However, the samples would still need to be analysed in wet-lab, as to give the manufacturer the statistical information of variables belonging to each NIRS class. This would possibly yield classes less homogeneous concerning wet-lab attributes uncertainty but could form classes which comprise more latent information, than the ones performed in this study. Additionally, different classification model algorithms such as the more traditionally used soft independent modelling of class analogy (SIMCA) (Bevilacqua et al. 2013; Gemperline et al. 1989); or the recent trend of ensemble methods (e.g., Random Forest) have to be explored, as they might yield better performances (Lee et al. 2013; Mevik et al. 2004; Nawar and Mouazen 2017).

It is shown that NIRS can also characterize the citrus peel raw materials with PLS-R predictive models of three variables of u_{peel} (IV_0, $\%DE_0$ and C^0_{pectin}) and providing a reasonable estimate for $\%Yield$ (Table 4.4). For an unbiased estimator model, as in the PLS-R example, the root mean square error of prediction (RMSEP) is equivalent to the standard deviation. The model predictions for$\%Yield$ exhibit a poorer correlation, understandably so since it is a material attribute that is less explicitly related to direct correlation in the spectra, and depends more on the full

extent of a test extraction. However, a RMSEP=1.57% is still within the uncertainty range of the other methods and this approach allows having a quick estimate, not reliant on wet-lab analysis of the sample in question. Additionally, these predictions provide a more accurate central tendency: a sample which would be in the extreme of a group statistic, thus with the group mean being a bad estimation of its true value, is individually predicted by the PLS-R model.

It should be stressed that the comparison of approaches in this study is made under the premise of negligible noise in the reference values compared with both the class uncertainty distribution and the prediction uncertainty. However, this is not always the case, especially when dealing with biologically derived raw material where the biotic noise may severely influence the accuracy of the reference method. This is an additional point in favor of using NIRS, as it has been shown that predictions for a group of samples can be closer to their true values than the set of lab analysis for this same group of samples (DiFoggio 1995), In fact, the RMSEP calculated in this study (Table 4.4) are in reality the apparent RMSEP, dependent on both the errors in the lab values and the inherent model errors and can be a pessimistic expectation of prediction uncertainty. An effective correction for the reference error component leads to $RMSEP_{corrected} = \sqrt{\left[RMSEP^2_{apparent} - \hat{\sigma}^2_{error}\right]}$ where $\hat{\sigma}^2_{error}$ is the estimate of variance of the reference method (Næs et al. 2004). This error estimate can be computed with the full analysis of the repeatability and reproducibility (e.g., Gage R&R) of the reference method (Deshpande et al. 2014).

When applicable, PAT tools should be favored as the in-process benefits are manifold. The use of NIRS raw material identification has the significant advantage of enabling at-line analysis directly at the reception in the warehouse or the feed-inlet of the tanks. Another aspect to be considered is that raw material may suffer in-lot variation and wet-lab analytical methods are performed on a minute amount. This can lead to production quality drifts if the lot in question is used continuously and the estimates on the critical material attributes are not adapted to this in-lot variations. A spectroscopic approach would be a low-cost and straightforward strategy to screen this variation continuously. Overall, PAT tools offer significant improvements in speed of analysis and resource wasting, enabling faster decision-making. The use of PAT does not motivate a complete elimination of the wet-lab analysis set-up. The calibration models need to be continuously updated, and the manufacturer needs to ensure new suppliers are within the previously established design space for raw materials, but it alleviates the quality control laboratory from the production optimization support task and allows for fewer tests.

4.4 Concluding remarks

The citrus peel impact has been assessed and deemed to be of critical concern on the pectin quality and production yield. Thus as different approaches to determining critical material attributes were explored in order to have a way which enables

our operational strategy. The path from a heuristic-based to a PAT-based operation is presented, highlighting the key differences regarding time-expenditure (in development and in-operation) together with the information each approach provides the manufacturer. Characterization of dried citrus peels through multivariate data analysis was performed and illustrated the successive developmental nature of the different approaches. An investigation on how the quantitative performance of chemometric near-infrared spectroscopy prediction models compare with dataset class statistics (based on expert-knowledge or clustering algorithm classes) was assessed. The non-invasive spectroscopic method has been proven to have the potential to characterize pectin extraction raw material with minimal sample preparation. The study shows the potential benefits of opting for PAT-based approaches on the early stage of bioprocessing plants when compared with industry standards of "rule-of-thumb" operation.

CHAPTER 5

Model Analysis

In this chapter particular focus is given to the effect of the critical process parameters in the solid-liquid extraction profiles in pectin production. The relationship between these, the material attributes and the key performance indicators is established. Section 5.1 introduces the methods used both in the empirical analysis of the data as well as the model-based procedures. Section 5.2 explores the kinetics of the extraction and the influence of the critical process parameters in the outputs. Section 5.3 presents the results of the analysis to the model developed by Andersen et al. (2017), highlighting the raw material variability impact.[†]

[†]This chapter includes sections which are based on parts of the published article:
Caroço, R.F., Kim, B., Santacoloma, P., Abildskov, J., Lee, J.H., Huusom, J.K. (2019). "Analysis and Model-based Optimization of a Pectin Extraction Process". In: *Journal of Food Engineering* 244, pp. 159-169

5.1 Methods

5.1.1 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) is performed to study the significance of independent variables on C_{pectin}, IV, and $\%DE$ extraction profiles and used to determine the regression coefficients of a statistical extraction model in the form presented in equation (2.1). The statistical significance of the model parameters are assessed for their F-value at a probability (p-value) of 0.001, 0.01 or 0.05. The model accuracy was also evaluated in terms of the R^2 value, and standard deviation. The F-value is the ratio of mean square error to the pure error, the p-value indicates the significance of the variables, whilst R^2 is the degree of fit. This analysis is carried out with the software JMP® 13 (SAS Institute Inc., USA) statistical software.

5.1.2 Model-based methods

The model developed by Andersen et al. (2017) is selected for analysis since it is adequate for full-scale operation optimization as it relates the critical process parameters (i.e., temperature, pH and raw material attributes) with the key performance indicators (i.e., pectin concentration in bulk (C_{pectin}), $\%DE$ and IV). Furthermore, the model describes the intra-particle solute diffusion and solubilization reaction, offering a kinetic process understanding which is be useful for process analysis. This model enables simulations varying these operational parameters, which is a key characteristic for development of optimization and operational strategies. The mechanisms and phenomena that are encompassed by the models are (in phenomenological order): Protopectin acidic hydrolysis, diffusion of pectin inside the peel, transfer of pectin through the boundary layer to the bulk, and degradation and de-esterification of pectin in bulk.

The modelling approach has the following assumptions:

- The peel is a matrix where the active component (pectin and protopectin) is dispersed homogeneously inside the matrix material.
- The existence of insoluble protopectin inside the peel, which is converted to pectin by acid hydrolysis.
- Water penetrates immediately into the inside of the peel and leads to release of pectin.
- Viscosity effects on diffusion and mass transfer are neglected.
- The peel has a constant geometry: Quadratic flake shape with a dimension of 1 cm and a thickness of 2 mm. The flake is planar and thin, approximated to slab geometry in one dimension.
- Diffusion is described by Fick's law.

Moreover, to enable the calculation of initial concentration of protopectin, it is considered that all protopectin is hydrolysed to pectin in the acid extraction and released from the peel at final time (t_f).

Table 5.1 shows the model, which includes the mass transport and reaction phenomena incorporating the influence of pH and temperature of the process by employing modified Arrhenius type functions for all the reaction rates (i.e., hydrolysis, degradation and de-esterification) as well as for the diffusion coefficient.

It also provides the means for a kinetic process understanding via the model analysis in this chapter. According to the authors, the model parameters were estimated using experimental data for C_{pectin}, IV, and $\%DE$ obtained from extractions performed at a pilot scale plant. The Arrhenius functions are centred on a reference temperature to avoid high correlation between the pre-exponential factors and the activation energies. The pre-exponential factor is expressed as a linear function of the proton concentration in the solution. The estimation involve four parameter sets θ^i, where $i = \{1, 2, 3, 4\}$ is the parameter set index. Moreover, it was performed in a two-step procedure:

Step 1.

Experimental design:

C_{pectin} profile data from three water extractions (no added acid, meaning hydrolysis and degradation negligible) at three different temperatures (60°C, 70°C, and 80°C).

Parameter estimation:

Isolated estimation of parameter set $\theta^1 = \left[C^0_{pectin}, k_{masstranfer}, D_{0,centred},\ D_a\right]$, where C^0_{pectin} is the initial concentration of pectin in the peel, $k_{masstranfer}$ is the mass transport coefficient of pectin through the boundary layer to the bulk, and $D_{0,centred}$, D_a are parameters related to the diffusion of pectin inside the peel.

Step 2.

Experimental design:

C_{pectin}, IV and $\%DE$ profile data from a full factorial composite design of fourteen extractions at three temperatures (60°C, 70°C, and 80°C) and pH (1.5, 2.3, and 3.1).

Parameter estimation:

Estimation of $\theta^2 = [C^0_{protopectin}, E_{a,hydrolysis}, \alpha_{hydrolysis}, \beta_{hydrolysis}, E_{a,degradation}, \alpha_{degradation}, \beta_{degradation}]$, the $\%DE$ parameter set $\theta^3 = [\%DE_0,\ E_{a,de-esterification}, \alpha_{de-esterification}, \beta_{de-esterification}]$, and the IV set $\theta^4 = [IV_0,\ E_{a,IV}, \alpha_{IV}, \beta_{IV}]$.

From the θ^2 set, $C^0_{protopectin}$ is the initial concentration of protopectin in the peel; the remaining parameters are related to the kinetics of hydrolysis and degradation. The parameters $\%DE_0$ and IV_0 are the initial values of the degree of esterification and intrinsic viscosity in-peel, the remaining parameters are related to their kinetics over time of extraction.

Table 5.1: System of equations that describe the KPI of the pectin extraction process according to Andersen et al. (2017). The system is composed by partial differential equations (2.32 and 2.33), ordinary differential equations (2.34 and 2.35) and algebraic equations (2.36 and 2.37)

C_{pectin}	$\frac{\partial C_{protopectin,peel}(t,x)}{\partial t} = -k_{hydrolysis} \cdot C_{protopectin,peel}(t,x)$	(2.32)
	$\frac{\partial C_{pectin,peel}(t,x)}{\partial t} = D_{pectin} \frac{\partial^2 C_{pectin,peel}(t,x)}{\partial x^2} + k_{hydrolysis} \cdot C_{protopectin,peel}(t,x)$	(2.33)
	$\frac{\partial C_{pectin}(t,L)}{\partial t} = \frac{A_{total} \cdot k_{masstransfer}}{V_{total}} \cdot \left(C_{pectin,peel}(t,L) - C_{pectin}(t,L)\right) - k_{degradation} \cdot C_{pectin}(t,L)$	(2.34)
$\%DE$	$\frac{\partial C_{ester}(t,L)}{\partial t} = DE_0 \cdot f_{GA} \frac{A_{total} \cdot k_{masstransfer}}{V_{total}} \cdot \left(C_{pectin,peel}(t,L) - C_{pectin}(t,L)\right) - k_{de-esterification} \cdot C_{ester}(t,L)$	(2.35)
	$\%DE(t) = C_{ester}(t) / \left(f_{GA} \cdot \left(C_{pectin}(t) + C_{degraded\ pectin}(t)\right)\right) \times 100$	(2.36)
IV	$IV(t) = IV_0 \cdot \exp(-k_{IV} \cdot t)$	(2.37)

The model analyses are performed following the methodology described by Sin and K. Gernaey (2016). This approach has previously been used in other bio-industrial case-studies, where a model is analysed and posteriorly applied Prunescu and Sin (2013). From the analysis, we obtain a diagnosis of model reliability and guidance to adequate parameter estimation. Impact of the raw material uncertainty can also be investigated using the re-tuned model. The sequence of analysis is summarized in Table 5.2. The analysis on step 2 has been already tackled in a previous chapter (Chapter 4), while the rest is presented in this subsection.

5.1.2.1 Raw material variability

One feature of the model is entailing parameters considered as critical material attributes (raw material specific inputs) $u_{peel} = \left[C^0_{pectin},\ C^0_{protopectin}, IV_0,\ \%DE_0\right]$. Andersen et al. (2017) estimated u_{peel} together with the rest of the parameters. The data used for estimation originated from pilot extractions of a single type of peel (lime).

From the years of manufacturing with these fruits, it is known that different fruits are typically used to produce pectins with different specifications (May, 1997). It is then expected that the different fruit peel samples can be grouped reasonably well by fruit type when considering the u_{peel} variables. The fruit groups have different mean values, for the four u_{peel} variables. However, assuming the distributions of these variables for the different fruits it was seen in Chapter 4 that some overlap exists meaning that certain fruits (i.e., $\%Yield$ distributions for lime and orange) have similar quality parameters. Nevertheless, the consideration of grouping peels

Table 5.2: Sequence for model analyses

#	Step description	Output
1	Consider parameter subsets from the literature (Andersen et al., 2017)	θ^i $i = \{1, 2, 3, 4\}$ for the different subsets
2	Identify raw material parameters and quantify uncertainty	u_{peel} Mean (μ) and standard deviation (σ)
	Sensitivity Analysis	
3	List parameter significance ranking	$\delta^{msqr}_{i,j,k}$
4	Identify identifiable parameter subsets	γ_L
	New Parameter Estimation	
5	Identification of parameters	θ_{new}
	Uncertainty Analysis	
6	Standard Monte Carlo analysis to quantify uncertainty in the outputs	5^{th}-95^{th} percentiles

Table 5.3: Mean and standard deviation of measured raw material variables for three citrus fruits

	IV_0 (dl/g)		$\%DE_0$		C^0_{pectin} (g/l)		$\%Yield$	
	μ	σ	μ	σ	μ	σ	μ	σ
Lime	7.90	0.31	73.97	1.38	1.72	0.64	26.22	2.07
Lemon	8.37	0.97	75.24	2.16	1.72	0.49	23.69	1.57
Orange	6.11	0.57	69.84	2.83	1.07	0.39	25.97	0.95

by fruit results in close to normal distributions with reasonable standard deviations. The grouping by fruit type as given in Table 5.3, which is compiled from the raw material assessment in Chapter 4, will be considered for the further analysis.

5.1.2.2 Sensitivity analysis

Local sensitivity analysis relies on examining the sensitivity functions of different model outputs to the various model parameters, typically those corresponding to the measured variables. The analysis adds greatly to process understanding as it quantifies parameter-output relationship for a given model structure (Brun et al. 2002; Sin et al. 2010; Sin and P. A. Vanrolleghem 2007). The first part of the analysis is conducted by computing the partial derivatives of the model outputs with respect to each parameter in a given set of parameters. This is termed absolute sensitivity functions:

$$s^{absolute}_{i,j,k} = \frac{\partial y_j}{\partial \theta^i_k} \tag{5.1}$$

where i indicates the parameter set, j is the index of the model output $y = \big[C_{pectin}, \%DE, IV\big]$, k is the parameter index in the parameter set θ^i, and $\frac{\partial y_j}{\partial \theta^i_k}$ is the output variation with respect to a variation in parameter θ^i_k. Absolute sensitivity can be transformed into non-dimensional sensitivity values:

$$s^{ndim}_{i,j,k} = \frac{\partial y_j}{\partial \theta^i_k} \cdot \frac{\theta^i_k}{sc_j} \tag{5.2}$$

where sc_j is a scaling factor with the same physical dimension as y_j. In this analysis, the scaling factor is chosen as the mean value of the model output with index j. Parameter significance ranking is then obtained in terms of the delta mean square, $\delta^{msqr}_{i,j,k}$, as defined by Brun et al. (2001):

$$\delta^{msqr}_{i,j,k} = \sqrt{\frac{1}{N}\sum_{1}^{N}\left(s^{ndim\,T}_{i,l,j}\, s^{ndim}_{i,l,j}\right)} \tag{5.3}$$

where N is the number of samples. All parameters in the different θ sets are ranked, according to the $\delta_{i,j,k}^{msqr}$ for each model output. The $\delta_{i,j,k}^{msqr}$ measure indicates the importance of the individual parameters in the evaluated subset in terms of their impact on the output. The greater the $\delta_{i,j,k}^{msqr}$ value, the higher the importance of the parameter. If a given parameter does not significantly affect the model outputs, its true value cannot be estimated reliably from experimentally obtained output data, and it may have to be excluded from the estimated parameter set.

The second part of the sensitivity analysis provides insights into the near-linear dependency of the parameter sensitivity functions, for different parameter subsets within each θ^i set. For such, collinearity index, γ_L, is used. First, the Euclidean normalized non-dimensional sensitivities are calculated:

$$s_L^{norm} = \frac{s_{i,j,k}^{non-dimensional}}{\left\| s_{i,j,k}^{non-dimensional} \right\|} \tag{5.4}$$

λ_L, indicates the eigenvalues of the normalized sensitivity matrixes:

$$\lambda_L = eigen(s_L^{normT} s_L^{norm}) \tag{5.5}$$

where L indicates the index of the subsets of parameters. The collinearity index is calculated for each subset L, resulting from all combinations of parameters within θ^i.

$$\gamma_L = \frac{1}{\sqrt{min\lambda_L}} \tag{5.6}$$

The index γ_L will tend to infinity if the parameters in the subset L are linearly dependent. Conversely, for independent subsets the index will approach $\gamma_L = 1$, which is desirable for identifiability of the parameters. It has been reported that an empirical threshold within 5-20 is typically used to determine the identifiability of a subset. The definition of this value is reliant on prior experience of the model, and is thus an iterative process (Brun et al. 2002; Lencastre Fernandes et al. 2013).

5.1.2.3 Parameter re-estimation

The end-purpose of the model, in this case, is for it to be applied in an operational routine that can accept the variability of the raw material as a stochastic input. The parameter estimation procedure performed by Andersen et al. (2017) included the u_{peel} parameters, identified in the prior Chapter 4, in the parameter set that was identified. This estimation has used data from a pilot extraction of a single type of peel (lime), which implies that the estimated peel parameters are unique to that given peel. Moreover, the estimated kinetic parameters are as well related to other parameters such as the agitation, solvent type and peel-to-solvent ratio used at the experimental conditions, since these factors are not modelled explicitly. To have parameter estimates that are independent of the peel type, the u_{peel} vector is taken out of the estimation. Instead, we obtain these values experimentally. Both concentration

values are acquired with lab-scale extractions for the specific lime peel used by Andersen et al. (2017). The values obtained, $C^0_{protopectin} = 175$ kg/m^3 and $C^0_{pectin} = 108$ kg/m^3, differ slightly from the ones estimated, mostly in the initial amount of pectin, which was $\hat{C}^0_{pectin} = 77.3$ kg/m^3. On the other hand, the initial IV and $\%DE$ values are chosen from the lime group distribution, IV_0(lime group mean) = 7.90 dl/g and $\%DE_0$(lime lower 95% bound) = 71.9%, for the best estimation results (see Table 5.3).

5.1.2.4 Uncertainty Analysis

The uncertainty analysis is performed by employing the standard Monte Carlo technique. This technique requires four main steps (Lencastre Fernandes et al. 2013; Omlin and Reichert 1999): (1) Input uncertainty definition; (2) Randomized sampling with correlation control, using the Latin hypercube algorithm; (3) Simulation with the sampled values; (4) Evaluation of the results.

5.2 Empirical and kinetic analysis

The extraction profiles in Figure 5.1 show the result of varying temperature and pH while maintaining the other operational parameters (i.e., raw material, solid-liquid ratio). In terms of concentration, the extraction temperature of 80°C showcases the highest yields, while the most acidic pH enabled an extraction of pectin to a higher extent (see transition form pH=3.1 to pH=2.3) and led to faster extraction rates. The improvement with the temperature is associated with a higher solubility which promotes the mass transfer of pectin. The pH effect (and thus availability of $[H^+]$) influences the extent of the hydrolysis conversion of protopectin to soluble pectin.

However, the effect of prolonged extraction time at the T=80°C and pH=1.5 denotes seemingly a degradation effect of pectin concentration. The yield increases with temperature, acidity and extraction time, but the length of the polymer also decreases since harsher condition cause the hydrolysis of the pectin galacturonan backbone. In terms of the CQA, the $\%DE$ and IV profiles reveal a similar tendency. From the different pH levels, the degradation effect on the methyl-ester groups ($\%DE$) is more evident rather than that of the decrease in the IV functional property. However, the temperature impact in the kinetics is steeper in IV.

5.2.1 Response surface methodology

Three second-order quadratic models are employed to determine the effect of the independent operational variables (temperature and pH) on the key process indicators at three different times $t_1 = 60$ min, $t_2 = 150$ min, $t_3 = 240$. The three levels mentioned before for temperature and pH were considered. All models were fitted by applying standard stepwise regression, adding and removing variables by stepwise

combination of both forward selection and backward elimination with a stopping rule of p-value threshold probability to enter = 0.1 and remove = 0.1). ANOVA is used to evaluate the adequacy of the models to characterize the extraction profiles. The significance of each coefficient is assessed through their F-test and p-values. A higher F-value and lower p-values implies that the term is more significant to the modeled response.

In a first iteration, the model is significant (F-value=95.96, p-value<0.0001). However, a lack of fit F-value of 6.44 (p-value=0.009) implies that the models lack of fit is significant. This can occur when there are missing terms in the model or if several, unusually large residuals result from fitting the model. From Figure 5.2 it can be seen that two data-points which have higher residuals. These are removed to assess

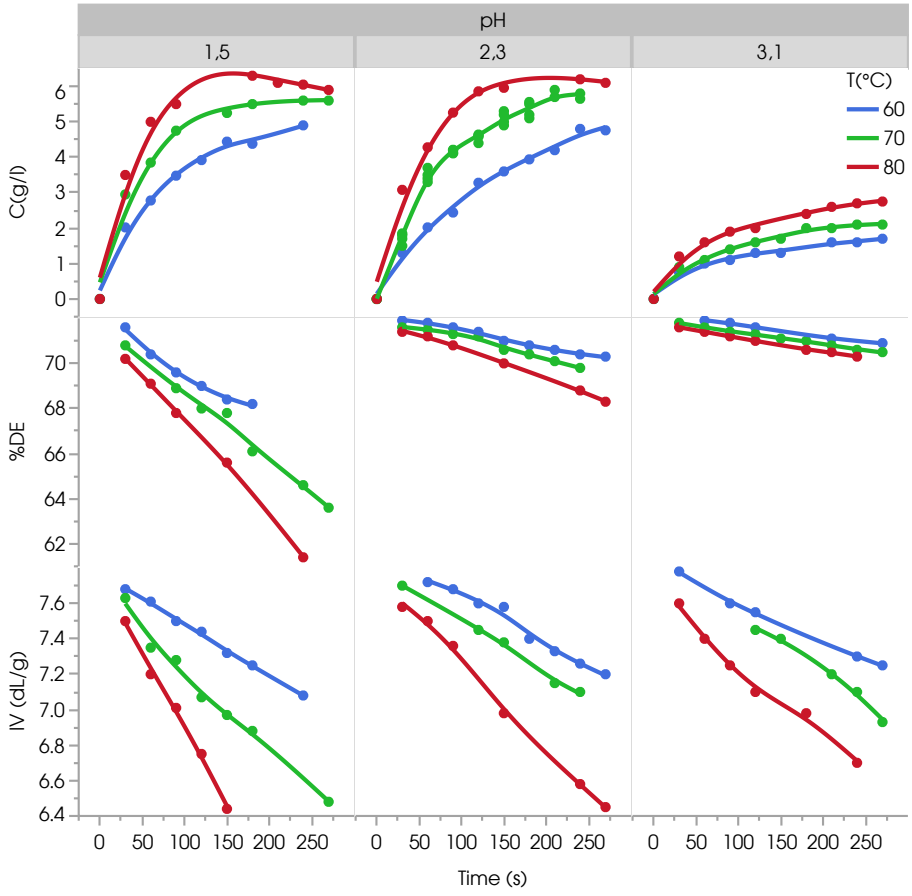

Figure 5.1: Data from the pilot-plant runs at DTU used in the estimation of parameters of the model developed by Andersen et al. (2017).

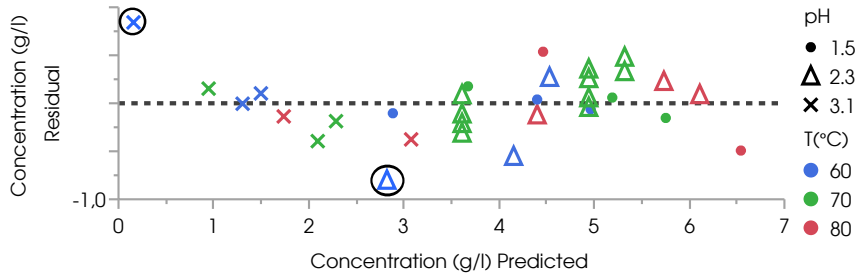

Figure 5.2: Residuals vs predicted Concentration plot. Colored and marked by the different temperatures and pH. The outliers are circled.

the interrelations of the variables with the concentration response in a way such that there is no lack of fit.

It can be seen from Table 5.4 that coefficient of determination (R^2) of the predicted model for concentration is 0.977 and p-value for Lack of Fit is 0.0649 (>0.5, non-significant). These provides indication that the model has relative good fit to the data. However, in Figure 5.3 the removed outliers are predicted (externally) and it

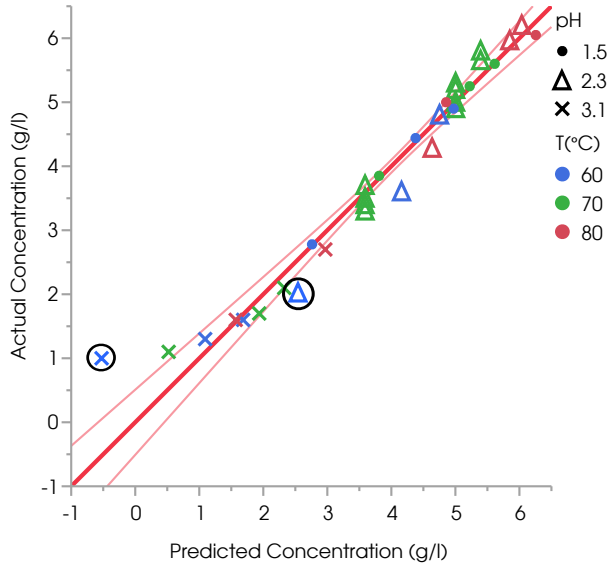

Figure 5.3: Predicted vs measured Concentration plot. Colored and marked by the different temperatures and pH. The red line is the 1:1 fit of the model with the 95% confidence interval. The circled data points are the excluded outliers.

Table 5.4: ANOVA result of model fit to concentration data

Source	Degree of Freedom	Sum of Squares	Mean Square	F-value	p-value
Model	6	69.663	11.611	163.990	<.0001
Residual	23	1.628	0.071	-	-
Total	29	71.292	-	-	-
Lack Of Fit	16	1.430	0.089	3.147	0.065
R^2	0.977				
$R^2_{adjusted}$	0.971				
RMSE	0.266				

can be seen that for the most extreme point in the selected range (T=60, pH=3.1 and time=60min) the model is unable to make a sound prediction. The other removed outlier (T=60, pH=2.3 and time=60min) improves in terms of its residual when comparing with Figure 5.2. It is observed in Table 5.5 that the variable with the biggest effect on concentration was pH, followed by time and quadratic form of pH. This is similar to what Gan and Latiff (2011) observed in their pectin extraction yield model, except in their model the linear and quadratic terms of time of extraction did not have significant effect. This is due to the fact that in their system, the extraction profiles would have already reached a plateau phase at their considered time range (2-4 hours), whilst this model captures the diffusion stage, as it can be seen in Figure 5.1 for the data obtained by Andersen et al. (2017).

For the development of the IV and $\%DE$ models it is not possible to conclude regarding their lack of fit. There is no lack of fit statistics if there are no replicates or if there are more unique design points than model coefficients. The former is the case in this analysis. For a process optimization based on response surface methodology a more careful design is necessary, with emphasis on replication. However, the interest at this point is to assess the influence of the different variables in the responses. Both

Table 5.5: Concentration regression model coefficients. Ordered by significance level.

Term	Estimate	Lower 95%	Upper 95%	p-value
pH	-1.936	-2.116	-1.755	<.0001
$Time$	0.010	0.009	0.0114	<.0001
pH^2	-2.225	-2.546	-1.904	<.0001
T	0.084	0.069	0.100	<.0001
$Time^2$	$-6.32{\cdot}10^{-5}$	$-9.08{\cdot}10^{-5}$	$-3.56{\cdot}10^{-5}$	<.0001
Intercept	2.041	0.874	3.208	0.001
$Time \cdot T$	-0.0002	-0.0004	$-2.18{\cdot}10^{-}5$	0.032

models $\%DE$ ($R^2 = 0.976$, RMSE=0.383) and IV ($R^2 = 0.974$, RMSE=0.072) were significant (p-value<0.0001) show a good correlation to the predictors that showed significant effect (p-value<0.05). Just as for concentration, the most significant variables affecting $\%DE$ are the linear pH and time terms. For IV the linear time and temperature effects are more significant. Both include quadratic effect of pH as a very significant term.

5.2.2 Peleg model

The sorption model introduced by Peleg (1988) can be used to explain the extraction curves of biological materials from plant sources because of its shape similarity. An initial estimate of the K_1 and K_2 parameters is made by linearizing equation (2.2) to:

$$\frac{1}{C_{pectin}} = K_1 \cdot \frac{1}{t} + K_2$$

This is performed for the center point of the dataset (T=70° C and pH=2.3). The resulting parameters are taken as initial guesses in a non-linear least squares estimation made with the MATLAB *lsqnonlin* function. The model fit to the data can be seen in Figure 5.4 with $K_1 = 723 \pm 8.45\%$ s and $K_2 = 0.116 \pm 6.34\%$ l/g.

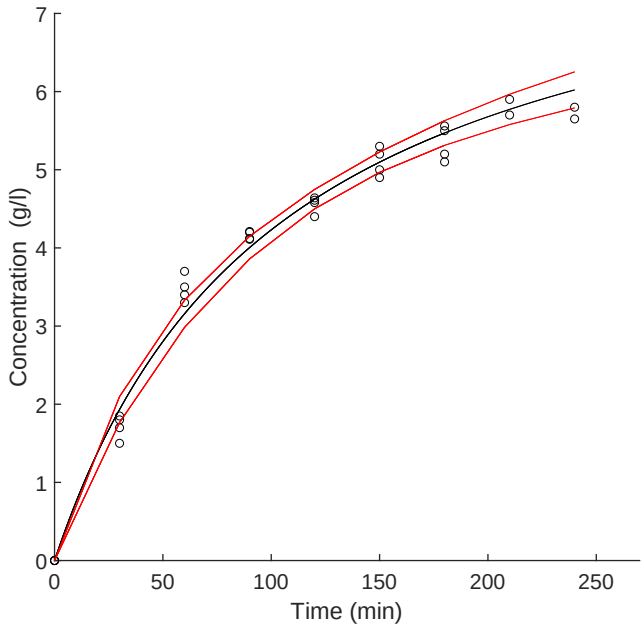

Figure 5.4: Peleg model fit to data from T=70° C and pH=2.3 experimental runs. $R^2 = 0.987$ and RMSE=0.2122. The red lines represent the 95% confidence interval.

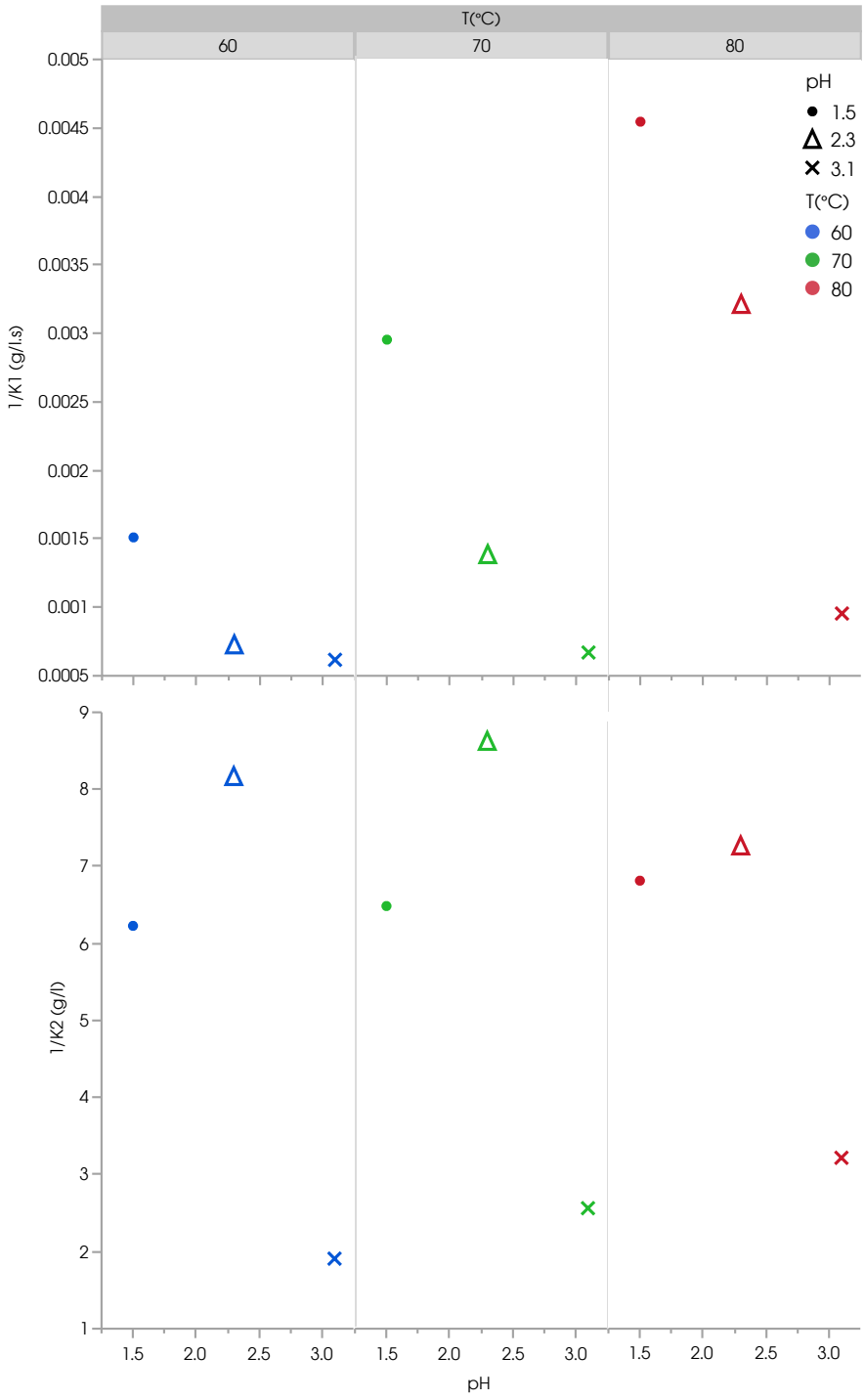

Figure 5.5: Effects of temperature and pH on the initial extraction rate, $1/K_1$, and the extraction extent $1/K_2$.

The process is modeled by these parameters in two stages, i.e. at the very beginning ($t = t_0$) it is a first order extraction rate constant ($B = 1/K_1$) which controls the process while in the latter phase of the process the order tend to zero with the K_2 constant relating to concentration at equilibrium ($C(t)|_{t\to\infty} = 1/K_2$).

The effect of temperature and pH on this parameters is demonstrated in Figure 5.5. The values of extraction rate constant show a clear tendency to increase with the increasing extraction temperatures and lower pH values. It is also seen that the increase in rate by variation of temperature is more accentuated at a lower pH. The influence of pH on extraction rate is also stronger at the higher temperatures up until the difference from pH=2.3 to pH=1.5 at T=80°C.

The capacity term, $1/K_2$, shows that at constant pH there is a small effect by increasing temperature. The exception is the $1/K_2$(T=80°C, pH=2.3), which is due to this profile having data points at t=240 min and t=280 min which suggest that an equilibrium is reached at that plateau level, while for $1/K_2$(T=70°C, pH=2.3) the data implies that the extraction has yet to reach its full extent, projecting a higher equilibrium concentration (see Figure 5.1).

If one removes these latter points, the same pattern as in the other pH conditions is observed at pH=2.3. The capacity constant for $1/K_2$(T=70°C, pH=2.3) is probably overestimated due to a slower $1/K_1$ rate and data points at later times would be necessary to assess this. This discrepancy in the measured extraction extent and projections between different runs could also explain why at fixed temperature T=70°C we observe $1/K_2$(pH=2.3) $> 1/K_2$(pH=1.5). However, for T=80°C this is also verified and both have apparently reach equilibrium, implying that another phenomena such as degradation could be the cause.

The temperature Arrhenius relationship of the extraction rate constant $1/K_1$ has been previously assessed (Buci-Koji et al. 2006). The Andersen et al. (2017) experiments follow this relationship as it is seen in Figure 5.6 A) for pH=2.3. From this linear fit, the slope provides us with a guess an estimate of the activation energy (E_a), whilst the intercept provides the natural logarithm of the frequency factor ($\ln(K_0)$) through the following equation:

$$\ln(1/K_1) = \ln(K_0) - \frac{1}{RT} \cdot E_a$$

To test if the Peleg model is appropriate for the description of the extraction process at different temperatures in this system, the estimates from the linear fit are then used as initial guesses in non-linear least squares estimation with temperature centering (at reference temperature T=70° C):

$$1/K_1 = B = K_0 \cdot \exp\Big(-\frac{E_a}{R \cdot T_{ref}}\Big) \cdot \exp\left(-\frac{E_a}{R} \cdot \Big(\frac{1}{T} - \frac{1}{T_{ref}}\Big)\right)$$

The K_2 is assumed constant for the fixed pH=2.3 (see Figure 5.5) and the mean value for the different temperature is used as initial guess for the estimation. The parameters obtained are shown in Table 5.6, together with their correlation matrix.

Table 5.6: Estimated parameters, standard deviation and correlation matrix for Arrhenius transformation of K_1 parameter from T=60° C, 70° C, 80° C at pH=2.3 extraction data.

	$\hat{\theta}$	σ	E_a	K_0	K_2
E_a (J mol^{-1})	65149	3474	1	0.4188	0.3875
K_0 (s^{-1})	0.00158	$6.26 \cdot 10^{-5}$		1	0.8736
K_2 (l/g)	0.13041	0.00293			1

As expected the K_2 is highly correlated with pre-exponential factor K_0, but no correlation between K_0 and E_a is observed due to temperature centering. Despite the correlation, indicating that the estimation will be very sensitive to the initial guesses for correlated parameters, the uncertainty is relatively low.

The model exhibits a good fit to data, $R^2 = 0.985$ and RMSE=0.227, with the caveat of over-predicting the t=240 min and t=280 min at T=80° C due to reasons mentioned before, and this being a joint estimation of all temperature data at pH=2.3 (see Figure 5.6 B)). The Peleg model exhibits to be a good representation for the specifics of this system.

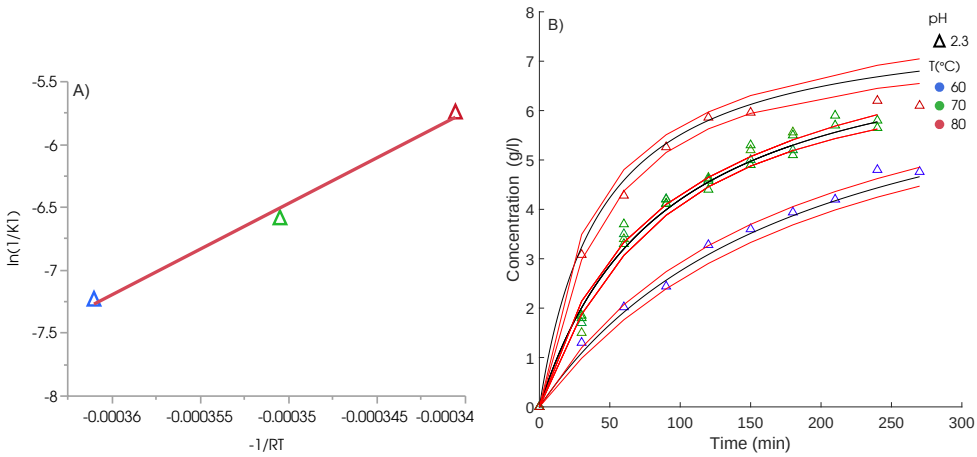

Figure 5.6: A) Arrhenius relationship plot for $1/K_1$ rate for fixed pH=2.3 B) Peleg model fit to data from T=60° C, 70° C, 80° C at pH=2.3. $R^2 = 0.9853$ and RMSE=0.2272. The red lines represent the 95% confidence interval.

5.2.3 Fick's effective diffusion

The effective diffusion of the different conditions can be calculated, considering a flake geometry, by adaptation of equation (2.14):

$$ln\left(\frac{c_\infty - c}{c_\infty}\right) = -0.21 - \frac{2.467 D_{eff}}{a^2} \cdot t$$

where a is the half thickness of the peel flakes ($a = 1$ mm), and the concentration of solute in solvent after infinite time (c_∞, at equilibrium) is given by the previously determined $1/K_2$ constants from the Peleg model.

The different D_{eff} are calculated by linear fit to the data ranging t=30-150 min and are shown in Table 5.7. It is observed that the effective diffusion increases with temperature at standard pH, expect at pH=3.1 where the value for D_{eff} remains constant (%95 confidence intervals coincide).

These results, combined with Figure 5.5, suggest that at pH=3.1 the extraction does not reach the same equilibrium concentration level as the other pH levels due to potentially not fully hydrolyzing the existing protopectin. The extraction kinetic is then not changed significantly with the increasing temperature. At lower pH, the equilibrium concentrations will be much higher, explaining why D_{eff}(T=60°C, pH=2.3) and (T=70°C, pH=2.3) exhibit slower rates than D_{eff}(pH=3.1).

Table 5.7: Estimated effective diffusion parameter and normalized 95% confidence intervals for the different conditions.

D_{eff} (m^2/s)		T (°C)		
		60	70	80
pH	1.5	$4.54 \cdot 10^{-11} \pm 3.17\%$	$7.13 \cdot 10^{-11} \pm 13.18\%$	$1.13 \cdot 10^{-10} \pm 9.59\%$
	2.3	$1.48 \cdot 10^{-11} \pm 28.36\%$	$3.10 \cdot 10^{-11} \pm 6.54\%$	$7.46 \cdot 10^{-11} \pm 8.12\%$
	3.1	$5.28 \cdot 10^{-11} \pm 8.81\%$	$4.14 \cdot 10^{-11} \pm 5.26\%$	$4.73 \cdot 10^{-11} \pm 11.80\%$

5.2.4 Consecutive reaction model

The data in Figure 5.1 is fitted with the Panchev et al. (1989) model (Eq. 2.30 and 2.31) by a non-linear regression, based on the *lsqnonlin* method using MATLAB software. The initial protopectin content is assumed from the estimated Peleg parameter $1/K_2^{max} = 8.62$ g/l which is obtained at the extraction conditions T=70°C and pH=2.3. The parameters K_{hyd} and K_{deg} are estimated for the different conditions. The fit for the center of this dataset can be seen in Figure 5.7, and there is a slight improvement when compared with the Peleg model (see Figure 5.4).

The correlation coefficient (R^2) indicates good fits for all the extractions (Table 5.8). It is seen that for constant pH, K_{hyd} increases and K_{deg} generally decreases with the increasing temperatures. This corroborates with the results obtained by Panchev et al. (1989) and Sebaoui et al. (2017). However, due to overlapping confidence intervals, K_{deg} can be said to be constant or without significant variability throughout pH=1.5-2.3 and T=70 – 80° C), at pH=3.1 at a higher value. This was not assessed by the authors and it can better visualized in Figure 5.8.

This higher value at pH=3.1 is un-intuitive. However, it was assessed by the Peleg model that the equilibrium capacity ($1/K_2$) for pH=3.1 is much lower. So, for the adopted $1/K_2^{max}$ the model had to adapt the degradation term (increase above the hydrolysis rate) in order to portray that the kinetics of never reaching that plateau.

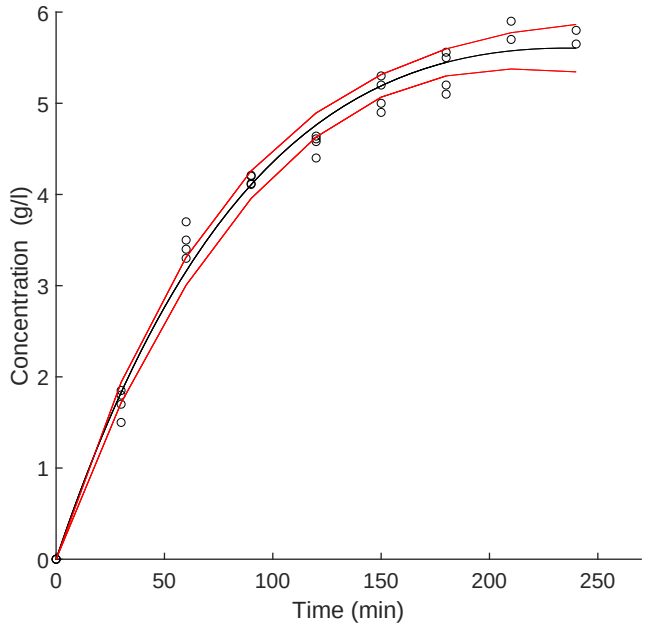

Figure 5.7: Panchev model fit to data from T=70° C and pH=2.3 experimental runs. $R^2 = 0.988$ and RMSE=0.2036. The red lines represent the 95% confidence interval.

Table 5.8: Estimated parameters for the consecutive reaction model, normalized 95% confidence intervals and correlation coefficient (R^2) for the different conditions.

$K_{hyd} \cdot 10^4$ (s^{-1})		T (°C)					
$K_{deg} \cdot 10^5$ (s^{-1})		60		70		80	
pH	1.5	$1.18 \pm 13.1\%$ $4.68 \pm 27.4\%$	0.98	$1.83 \pm 14.7\%$ $3.78 \pm 20.6\%$	0.97	$2.61 \pm 10\%$ $3.18 \pm 12.2\%$	0.98
	2.3	$0.72 \pm 9.37\%$ $2.38 \pm 36.5\%$	0.99	$1.37 \pm 4.62\%$ $3.04 \pm 13.5\%$	0.99	$2.1 \pm 8.66\%$ $2.81 \pm 16.2\%$	0.99
	3.1	$0.38 \pm 19.7\%$ $11.8 \pm 26.5\%$	0.94	$0.44 \pm 14.1\%$ $9.59 \pm 20.8\%$	0.97	$0.61 \pm 17.6\%$ $8.6 \pm 23.9\%$	0.96

The same effect can be seen in the work of Sebaoui et al. (2017) at pH=3.1.

The values of our hydrolysis parameters (K_{hyd}(80°C, 1.5)= $1.57 \cdot 10^{-2}$ min^{-1}) are between the lower values obtained by Pagán and Ibarz (1999) (K_{hyd}(80°C, 1.54)= $0.77 \cdot 10^{-2}$ min^{-1}) and the higher values of Sebaoui et al. (2017) (K_{hyd}(80°C, 1.5)= $5.04 \cdot 10^{-2}$ min^{-1}) for similar temperature and pH conditions.

The temperature dependency of K_{hyd} can be expressed by an Arrhenius relationship. The values for the pre-exponential factor and activation (obtained by linear regression) are listed in Table 5.9. Both pH=1.5 and 2.3 showed a $R^2 > 0.9$, while pH=3.1 had a lower correlation coefficient of $R^2 = 0.62$. The activation energy are higher than the one reported by Cho and Hwang (2000) (E_a = 17.77 kJ/mol) for

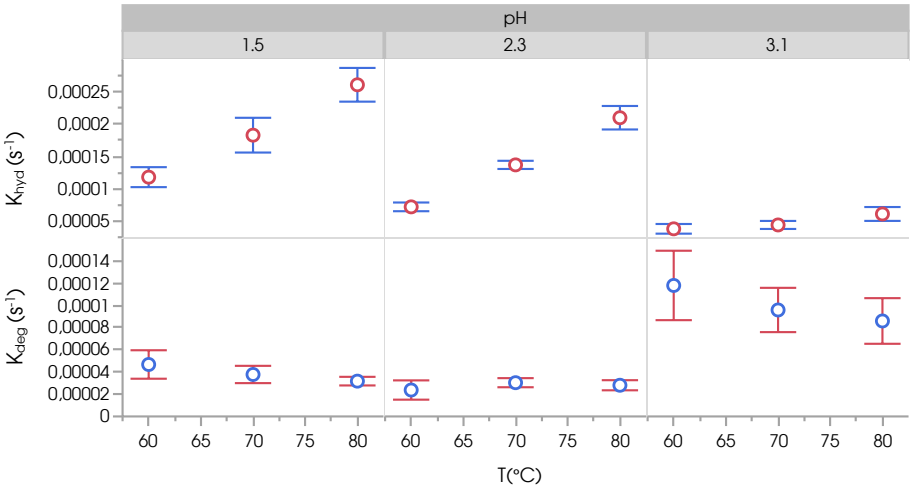

Figure 5.8: Effects of temperature and pH on the hydrolysis rate constant, K_{hyd}, and the degradation rate constant K_{deg}. The circles are the parameter estimates, surrounded by their 95% confidence interval.

apple pomace and Sebaoui et al. (2017) ($E_a(1.5) = 10.16$ kJ/mol) for lemon peel. This explains the smaller hydrolysis rate of this study when compared to these.

The temperature dependency of K_{hyd} can be expressed by an Arrhenius relationship. The values for the pre-exponential factor and activation (obtained by linear regression) are listed in Table 5.9. Both pH=1.5 and 2.3 showed a $R^2 > 0.9$, while pH=3.1 had a lower correlation coefficient of $R^2 = 0.62$. The activation energy are higher than the one reported by Cho and Hwang (2000) ($E_a = 17.77$ kJ/mol) for apple pomace and Sebaoui et al. (2017) ($E_a(1.5) = 10.16$ kJ/mol) for lemon peel. This explains the smaller hydrolysis rate of this study when compared to these.

It is evident from Figure 5.8 and 5.9 that the extractions at pH=2.3 and 1.5 follow a more similar extraction dynamic than the extractions at pH=3.1. To encompass the acid-catalyzed hydrolysis of protopectin to pectin at these conditions the reaction rate constants can be represented as the following modified Arrhenius equation (Girisuta et al. 2007):

$$K_{hyd} = K_0^{hyd} \cdot \exp\left(-\frac{E_a^{hyd}}{R \cdot T_{ref}}\right) \cdot \exp\left(-\frac{E_a^{hyd}}{R} \cdot \left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right) \cdot [\mathrm{H}^+]^{\alpha} \qquad (5.7)$$

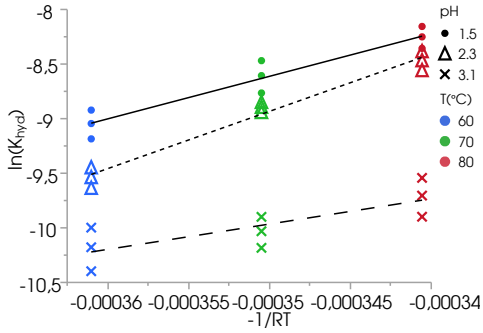

Figure 5.9: Arrhenius relationship plot for K_{hyd} at the different pH. The parameter estimations and their upper and lower 95% interval values are used.

Table 5.9: Arrhenius parameters estimated by linear regression for different pH.

pH	E_a^{hyd} (J mol^{-1})	k_0^{hyd} (s^{-1})
1.5	$3.90 \cdot 10^4$	$1.53 \cdot 10^2$
2.3	$5.25 \cdot 10^4$	$1.25 \cdot 10^4$
3.1	$2.32 \cdot 10^4$	$1.58 \cdot 10^{-1}$

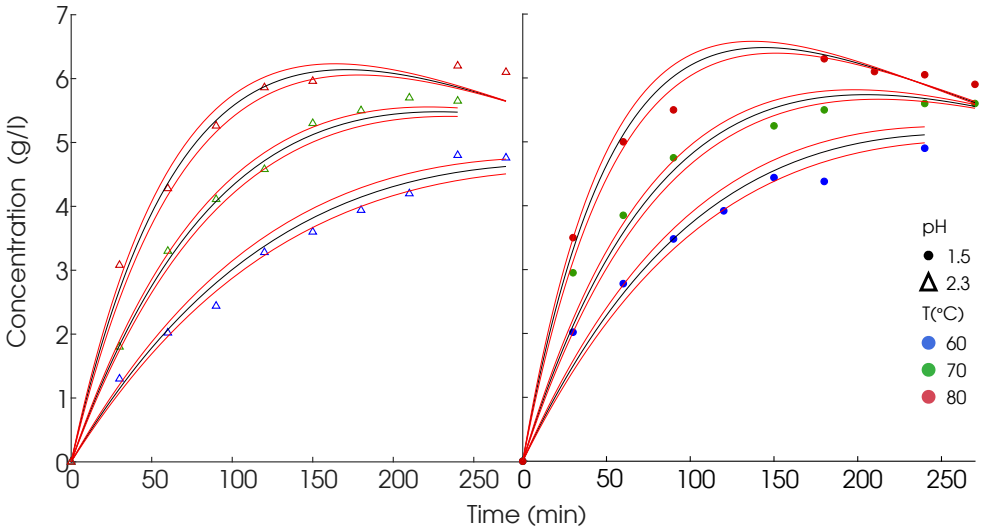

Figure 5.10: Panchev type model fit to data from T=60° C, 70° C, 80° C at pH=2.3 and 1.5. $R^2 = 0.98$ and RMSE=0.237. The red lines represent the 95% confidence interval.

where α is the reaction order in acid. The reaction rate constants is thus defined to include the combine effects of both temperature and acid-catalyst concentration. The E_a^{hyd} and k_0^{hyd} in Table 5.9 are taken as initial guesses for non-linear least squares estimation made with the MATLAB *lsqnonlin* function. The parameter $\bar{K}_{deg} = 3.31 \cdot 10^{-5}$ s^{-1} is assumed constant throughout the different conditions (see Fig. 5.8) and is not estimated with the rest.

A good fit between the experimental data and the modelling results was obtained. A coefficient of correlation (total) $R^2 = 0.98$ and RMSE=0.237 is achieved for the parameters $E_a^{hyd} = 4.81 \cdot 10^4 \pm 7.5\%$ J mol^{-1}, $k_{0,centered}^{hyd} = 2.65 \cdot 10^{-4} \pm 14.9\%$ s^{-1} and $\alpha = 0.13 \pm 25.7\%$. The model is fitted slightly better to pH=2.3, however at both pH the model over-estimates the degradation effect at T=80°C.

5.3 Model-based analysis

5.3.1 Parameter identifiability and estimation

In Figure 5.11A, it is observed how the different parameters estimated with the water extraction data rank relatively to their $\delta_{i,j,k}^{msqr}$ measures. The output state C_{pectin} is most sensitive to C_{pectin}^0, and the effect of $k_{masstransfer}$ is negligible in comparison. Also, the diffusion of pectin inside the peel appears to be a more significant phenomenon than the mass transfer of pectin through the boundary layer to the bulk.

This appears to corroborate the findings of several researchers engaged in the solid-liquid extraction of natural products from plant materials that the rate-limiting step of the process is the diffusion of the dissolved solute within the solid (Chan et al. 2014; Hojnik et al. 2008; Pinelo et al. 2005; Seikova et al. 2004; Simeonov et al. 1999). The pre-exponential hydrolysis parameters, directly involved in the pH dependence term, are the most significant in the evaluation of the second parameter set (see Figure 5.11B). Once more, the raw material parameter, $C^0_{protopectin}$, ranks high. The degradation parameters, compared to those for the hydrolysis parameters, are not significant. These insights provide good support for the conclusion that degradation is not a dominant factor in comparison to the hydrolysis and diffusion, indicating the possibility of a model simplification. The sensitivity analyses of the outputs $\%DE$ and IV, (Figure 5.11C and Figure 5.11D), show once again that the feedstock parameters are highly significant for all outputs. Regarding the collinearity index, for this case study it is considered that any subset L with a collinearity index value above $\gamma_L > 15$, is poorly-identifiable or non-identifiable. For the set of estimated parameters with water extractions, θ^1, these values can be seen in Table 5.10. Any subset of parameters that include C^0_{pectin} and $D_{0,centred}$ together is not-identifiable. If the concentration of initial pectin in the peel is kept out of the estimation set, the diffusion parameters can be uniquely identified. The collinearity analysis is as well performed for the acid extractions concentration output, and the subset containing all the parameters yielded a γ_L of 21.74, which is well above the threshold. If the model is simplified to neglect the degradation effect and the $C^0_{protopectin}$ is left out of

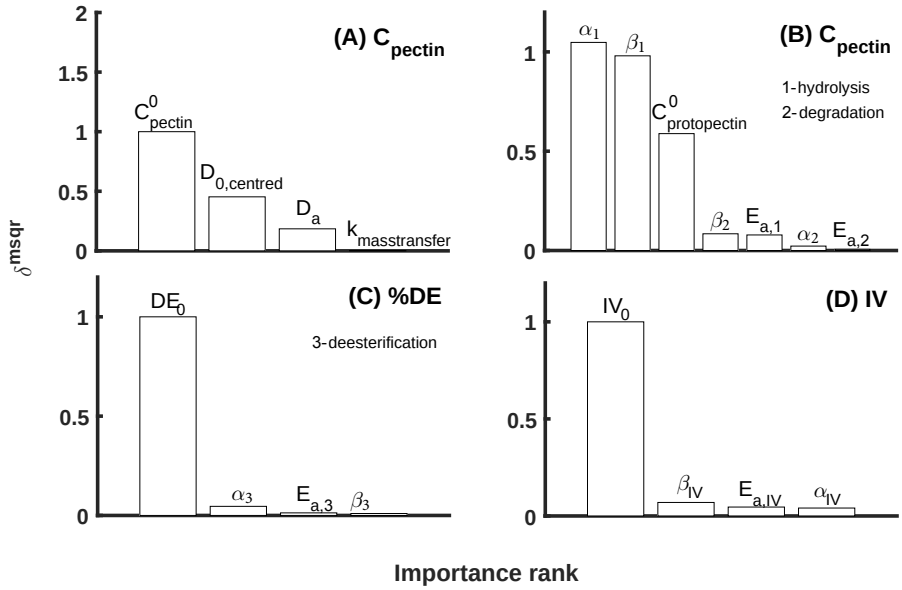

Figure 5.11: Parameter significance ranking for pectin concentration in the bulk solution, (A-θ^1 and B-θ^2), $\%DE$ (C-θ^3) and IV (D-θ^4), based of δ^{msqr} values

Table 5.10: Collinearity indexes for all combinations of D_a, $D_{0,centred}$, $k_{masstransfer}$ and C^0_{pectin}

L	Size	θ Combinations				γ_L
1	2	$k_{masstranfer}$	C^0_{pectin}			3.68
2	2	$D_{0,centred}$	C^0_{pectin}			11.78
3	2	$D_{0,centred}$	$k_{masstranfer}$			4.04
4	2	D_a	C^0_{pectin}			1.05
5	2	D_a	$k_{masstranfer}$			1.03
6	2	D_a	$D_{0,centred}$			1.08
7	3	D_a	$D_{0,centred}$	$k_{masstranfer}$		5.01
8	3	D_a	$D_{0,centred}$	C^0_{pectin}		12.33
9	3	D_a	$k_{masstranfer}$	C^0_{pectin}		4.00
10	3	$D_{0,centred}$	$k_{masstranfer}$	C^0_{pectin}		12.38
11	4	D_a	$D_{0,centred}$	$k_{masstranfer}$	C^0_{pectin}	18.13

the estimation, only the three hydrolysis parameters are left in the set. This results in a γ_L of 14.26, which is slightly below the threshold. For IV and $\%DE$, the values were below four thus, indicating that the respective parameters could be uniquely estimated.

The combined information from the $\delta^{msqr}_{i,j,k}$ values and the collinearity indexes indicate that $k_{masstranfer}$, $E_{a,hydrolysis}$ and the coefficients for degradation ($E_{a,degradation}$, $\alpha_{degradation}$, $\beta_{degradation}$) cannot be identified reliably as they do not significantly affect the model outputs. Hence, they are left out of the estimation as well as the u_{peel} parameters as described in subsection 5.1.2.3. The remaining parameters for estimation are:

$$\theta_{new} = [D_{0,centred}, D_a, \alpha_{hydrolysis}, \beta_{hydrolysis}, E_{a,de-esterification}, \alpha_{de-esterification}, \beta_{de-esterification}, E_{a,IV}, \alpha_{IV}, \beta_{IV}]$$

The new estimates are given in Table 5.11. There are slight differences between the new and the previous estimates. However, the orders of magnitude of the reaction rates are maintained. Particularly, the new diffusion coefficients are lower and have less uncertainty. These lesser values could be explained by the fact that when including the initial concentration of pectin in the peel in the estimation routine, the optimal numerical data fit value for $\hat{C}^0_{pectin} = 77.3$ kg/m^3 is lower than the experimentally obtained $C^0_{pectin} = 107.89$ kg/m^3. This ultimately led to an overestimation of the amount of pectin that was released by immediate diffusion, contributing for a larger diffusion coefficient.

The new estimate is also less uncertain; this improvement is obtained at the expense of the hydrolysis parameters. However, this trade-off is preferred and is necessary since the extraction of pectin is a diffusion-dominant process (Durán et al. 2015). The remaining rates are similar, due to the u_{peel} values determined experimentally

Table 5.11: Values of estimated parameters with relative 95% confidence interval. $D_{0,centred}$ is converted to non-centred parameters for comparison with the values from Andersen et al., 2017.

	$\theta^i{}_{new}$	θ^i(Andersen et al., 2017)
D_a(J/mol)	$4.81 \cdot 10^4 \pm 0.43\%$	$4.91 \cdot 10^4 \pm 16\%$
D_0(m^2/s)	$9.17 \cdot 10^{-4} \pm 5.54\%$	$1.7 \cdot 10^{-3} \pm 20\%$
$\alpha_{hydrolysis}$ (l/(mol s))	$3.25 \cdot 10^{11} \pm 21.7\%$	$3.2 \cdot 10^{11} \pm 7.5\%$
$\beta_{hydrolysis}$ (s^{-1})	$-2.6 \cdot 10^8 \pm 21.6\%$	$-2.4 \cdot 10^8 \pm 10.4\%$
$E_{a,de-esterification}$(J/mol)	$2.92 \cdot 10^4 \pm 8.1\%$	$3 \cdot 10^4 \pm 54\%$
$\alpha_{de-esterification}$(l/(mol s))	$8.46 \pm 2.6\%$	$11 \pm 1.8\%$
$\beta_{de-esterification}$(s^{-1})	$4.7 \cdot 10^{-2} \pm 6.9\%$	$5.01 \cdot 10^{-2} \pm 4.2\%$
$E_{a,IV}$(J/mol)	$4.63 \cdot 10^4 \pm 7.5\%$	$5.32 \cdot 10^4 \pm 6\%$
α_{IV}(l/(mol s))	$2.1 \cdot 10^3 \pm 10.7\%$	$2.2 \cdot 10^4 \pm 14.5\%$
β_{IV}(s^{-1})	$8.13 \cdot 10 \pm 4.4\%$	$7.93 \cdot 10^2 \pm 1.6\%$

being similar to the ones estimated by Andersen et al. (2017), with a notably smaller uncertainty for the estimate of the activation energy of de-esterification.

5.3.2 Monte-Carlo simulations

Raw material parameters are chosen as the source of uncertainty. These variables, u_{peel}, are obtained by the method described in the previous chapter 4. Peel samples are grouped per type of fruit, and the raw material uncertainty is defined for each specific fruit group (e.g., lime). The mean and standard deviations for these groups are considered**.** Despite using a same amount of acid for the extraction of all peels, the pH in lab extractions varies according to peel. These variations are related to the different buffer capacity of each peel (Sinclair and Eny 1947). Since IV and $\%DE$ profiles are a function of pH and temperature (constant at 70°C in this study), the pH is expected to contribute to the input uncertainties derived from the raw material. The distributions of these parameters are close to "normal" and adopted as such. The covariance of the parameters is calculated so that it can be incorporated into Latin Hypercube sampling with correlation control method, as proposed by Helton and Davis (2003). The simulations are performed for 1000 random samples, and the output results are summarized by using the mean, standard deviation and percentile calculations. The simulations are carried out at input conditions that mimic the lab-scale conditions, at which the degradation effects are negligible due to the shorter operational times (simulation time $t_f = 240$ min), and less abrasive mixing. These results are compared with the experimental data. As a result of the Monte Carlo simulations, the mean values of model outputs along with the 5^{th} and 95^{th} percentile of the distribution are presented in Figure 5.12. All fruits showed different, albeit slightly overlapping, distributions and standard deviations for the inputs, thus leading to the different output distributions from the Monte Carlo simulations.

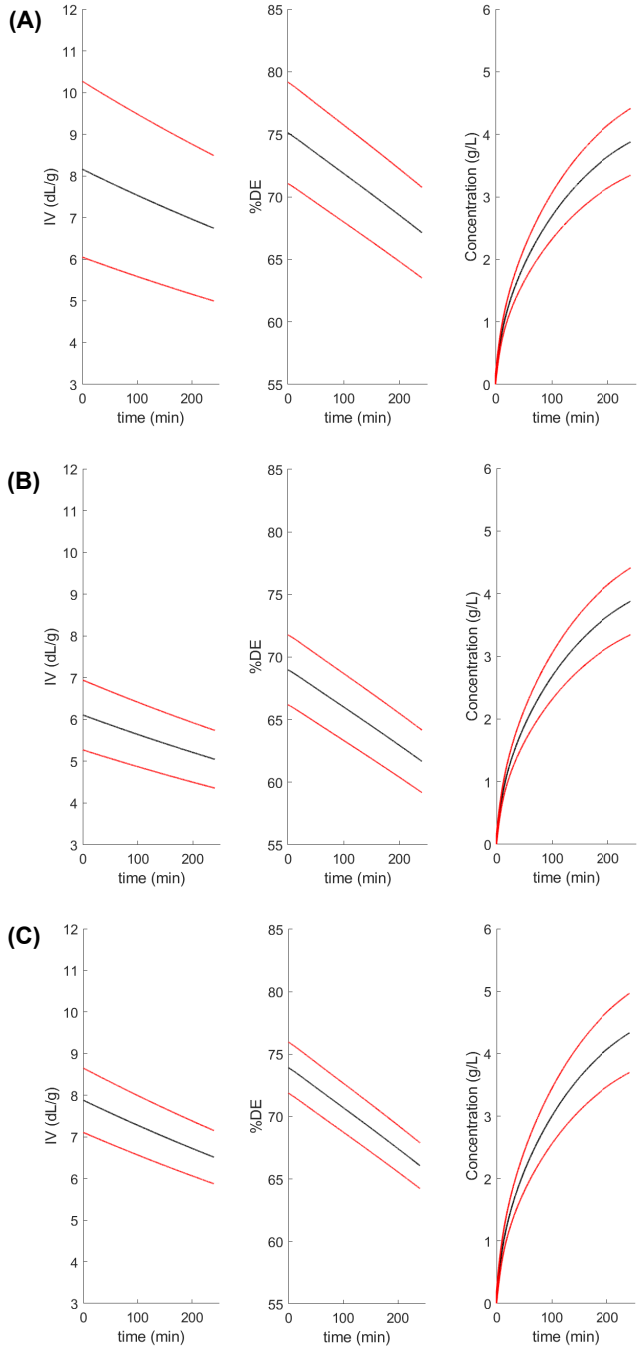

Figure 5.12: Model prediction uncertainties for IV, $\%DE$ and pectin concentration in the bulk solution of (A) lemon, (B) orange, (C) lime. The mean (solid line) and the 5^{th} and 95^{th} percentiles (dashed line).

Table 5.12: Simulation and experimental mean values and uncertainty of the outputs at t_f.

		Lemon	Orange	Lime
IV (Exp.)(dl/g)	μ	6.15	4.84	6.49
	σ	0.67	0.27	0.37
IV (MC)(dl/g)	μ	6.74	5.05	6.51
	σ	0.82	0.32	0.3
$\%DE$ (Exp.)	μ	67.99	64.41	67.10
	σ	1.51	0.76	0.82
$\%DE$ (MC)	μ	67.14	61.67	66.1
	σ	1.67	1.17	0.86
C_{pectin} (Exp.)(g/l)	μ	4.77	5.22	5.31
	σ	0.31	0.19	0.37
C_{pectin} (MC)(g/l)	μ	3.88	4.24	4.33
	σ	0.25	0.15	0.3

Orange has a lower mean, as well as a lower standard deviation, for the IV and $\%DE$ outputs when compared to lime and lemon. The latter fruit exhibits the widest uncertainty in the outputs, consequence from having largest input uncertainty distribution (see Table 5.3). This is mainly due to being the most utilized fruit in the raw material assessment, having a wider supplier portfolio. Also, there is notable overlap in the simulated output values between lemon and limes. As previously mentioned, there is overlap in the input for these fruits, meaning that lemons have certain u_{peel} variables with identical values to limes. Orange shows the least spread in the output, due to more homogeneous u_{peel} values. This goes against the statement that describes this fruit as being a production liability due to its fluctuations in quality (May 1990).

The model output with the largest uncertainty across all peels is IV, followed by C_{pectin}. This is due to the large variances in IV_0, C^0_{pectin}, and $C^0_{protopectin}$ in the peel, when compared with $\%DE_0$. Remarkably, the variance of IV of the raw material is far larger than the variance of $\%DE$ (Figure 5.12 and Table 5.12). Operating conditions should be carefully determined in consideration of the uncertainty in IV_0 when the final value of IV is constrained. The uncertainty in C_{pectin} varies over time (Figure 5.12) and increases as the hydrolysis and mass transfer of pectin occurs. This is caused by the variance in C^0_{pectin} and $C^0_{protopectin}$, whose contributions become more significant as the extraction proceeds. In Table 5.12, the mean of the final values of IV and $\%DE$ from the Monte Carlo simulations are similar to the experimental values. However, for all fruits, the mean of C_{pectin} from the Monte Carlo simulations are about 19% lower than the experimental values. This simulation mismatch is systematic, revealing that the problem resides in the dynamic process variables rather than the feedstock variability. The model parameter uncertainty, namely those in the hydrolysis parameters (Table 5.11), might be the source for this under-prediction, and

this issue deserves a further investigation. The confidence intervals range from ±4% to ±27%, relative to each output mean at t_f. This range shows that, when adopting fruit as a group for quality proximity, there is a considerable variation in the KPI. Peel specific analysis, by lab reference methods with possible coupling with spectroscopic process analytical technology, should be considered to reduce uncertainty in applying the model in production.

5.4 Concluding remarks

The process was evaluated in terms of its critical process parameters, through the analysis of models that mapped the relationship between these parameters and the process performance indicators. A series of kinetic models were analysed and showed good fit to the data. These models unravelled important insights into the interdependencies of the critical process parameters and the key performance indicators. However, it was seen that despite having a good fit they can lead to erroneous insights and due to their empirical nature, their use is limited to their system of design. As a necessary step prior to application, the pectin extraction dynamic model was analysed together with data from different types of raw materials, showing that the raw material-specific parameters (critical material attributes) are the most significant in terms of affecting the model outputs. An uncertainty analysis indicated an acceptable prediction performance considering the measured variables. The model output variance matched well with the measured variabilities of the extraction pectin concentration, intrinsic viscosity, and degree of esterification of the different fruits. This is an indication of the adequacy of the model as a tool for the optimization.

CHAPTER 6

Process Optimization

In this chapter particular focus is given to the procedure for the optimization of the critical process parameters given a certain raw material and CQA target. Section 6.1 introduces the optimization problem that is needed to solve and introduces a setting for an optimization. Section 6.2 solves the optimization problem deterministically, while Section 6.3 takes a robust approach to the problem, considering raw material attributes distributions.[†]

6.1 Methodology

A model-based optimization strategy design requires the identification of critical process parameters and definition of the objective function. The different optimization scenarios are defined according to quality specifications (IV and $\%DE$) of particular pectin-product applications which require different gelling abilities. The critical process parameters are then optimized for each scenario in a deterministic manner. According to the sensitivity and uncertainty analyses, initial conditions of the raw material u_{peel} significantly affect the product quality as well as the yield and have variations in mean (μ) and standard deviation (σ) across different types of fruits as assessed in Chapter 4 and further compiled in Table 5.3. Therefore, a robust optimization will finally be performed to decide the optimal operating conditions in consideration of the feedstock uncertainty and the required characteristics of extracted pectin.

[†]This chapter is based on parts from the published article:
Caroço, R.F., Kim, B., Santacoloma, P., Abildskov, J., Lee, J.H., Huusom, J.K. (2019). "Analysis and Model-based Optimization of a Pectin Extraction Process". In: *Journal of Food Engineering* 244, pp. 159-169

6.1.1 Optimization problem

In the pectin extraction process, temperature, pH, and time of extraction are the main operational degrees-of-freedom, affecting the important outputs of the process. A low pH promotes hydrolysis, de-polymerization and de-esterification leading to a high pectin yield along with low $\%DE$ and IV for a given extraction time (Masmoudi et al. 2008; Ziari et al. 2010). A high temperature has the same effect on the outputs as a low pH. Also, the extraction time is one of the important operating variables because a longer extraction time contributes significantly to the extent of the reaction for the de-polymerization and de-esterification (Masmoudi et al. 2008). Thus, a better pectin yield can be obtained at a higher temperature (T ≈80 °C), lower pH (pH≈1.5) and longer extraction time (t≈6 hours) but IV and $\%DE$are considerably reduced at these conditions. However, at the extreme condition (e.g., long extraction time with high temperature and acidity) it leads to an excessive degradation, resulting in a poor pectin recovery and so a lower yield. To achieve a high pectin yield with the desired characteristics, temperature, pH, and extraction time should be carefully controlled. Therefore, they are chosen as the optimization variables in this study.

A typical batch optimization problem is to maximize product yield or to minimize cost for a given batch time (Bonvin et al. 2001). For optimization of the extraction conditions, the objective function to be maximized is defined as pectin concentration in the bulk solution at the end of a batch. Moreover, requirements for the quality specifications of the extracted pectin are given as inequality constraints at the final time (t_f). The overall optimization problem is stated as below:

$$\begin{aligned} \underset{z}{\text{maximize}} \quad & J = C_{pectin}(t_f) \\ \text{subject to} \quad & F\left(\dot{x}, x, z, q, t\right) = 0,\ (i,j) \in \Omega, \\ & x\left(0\right) = x_0 \\ & z_l \leq z \leq z_u \\ & D\left(x\left(t_f\right)\right) \leq 0 \end{aligned} \tag{6.1}$$

Where J is the objective function of the pectin yield, which is to be maximized. The vector of operating variables z contains temperature, pH and extraction time t_f bounded by lower bounds z_l and upper bounds z_u. The state vector at the extraction time, $x\left(t_f\right)$, is derived from the simulation of the nonlinear system model F described in Table 5.1 (with the re-estimated parameters in Chapter 5) initialized with the initial states x_0. D represents the vector of constraints of the terminal states (i.e., $\%DE$ and IV) to achieve the desired pectin quality. Note that there are no path constraints on the states in this process. Other design variables q are set *a priori* based on the standard operating conditions at a commercial-scale production: the ratio between the amount of water (m_{water}) and peel (m_{peel}) is fixed at 17:1. The nonlinear constrained optimization is solved using *fmincon* function in MATLAB R2015b Optimization Toolbox.

6.1.2 Optimization scenario

Since HM pectins can be converted into LM pectins or LM amidated pectins through further de-esterification processes after the extraction (Ciriminna et al. 2015), different HM pectins are considered the desired products to which the operating conditions of the extraction are optimized. The model-based optimization of the pectin extraction process is performed for two product scenarios: pectins for productions of jam and jelly. For this, the desired characteristics of extracted pectin are translated as the terminal constraints in equations (6.1) based on the effects of $\%DE$ and IV on gel quality.

For the production of jam, a high rate of gel formation and moderate gel strength are required to avoid fruit floatation and ensure a uniform distribution of fruit particles. On the other hand, for the production of jelly, the pectin must set slowly with a high gel strength to allow air bubbles to escape (Walter, 1991). In other words, the pectin for jam (termed "jam case" hereafter) require high lower bound on $\%DE$ and a low upper limit on IV, and the pectin for jelly (termed "jelly case" hereafter) should have a low upper limit of $\%DE$ and a high lower bound on IV. The constraints for each of the scenarios are summarized in Table 6.1. For the base case, the optimization is performed without any constraint. Note that the constraints of the functional characteristics of the extracted pectin can be modified depending on the desired pectin quality.

6.2 Deterministic optimization

In initializing the dynamic simulation for the optimization, the initial concentrations of pectin and protopectin in the peels, and initial values of $\%DE$ and IV are set as the mean values of the three different types of fruits as reported in Table 5.3. Other design variables, m_{water} and m_{peel} are fixed as presented in section 6.1. Given the initial conditions, the deterministic model-based optimization is conducted by solving the nonlinear constrained optimization problem in equation (6.1) for three cases: base, jam and jelly cases (see Table 6.1). The results of the deterministic optimization are summarized in Table 6.2, Figure 6.1 and Figure 6.2. For all fruits the base case optimization gives similar optimal operating conditions when maximizing the final pectin concentration.

Table 6.1: Optimization scenarios for three cases: base, jam and jelly cases

Product type	Constraint of final IV(dl/g)	Constraint of final $\%DE$
Base	No constraint	No constraint
Jam	$5 \leq\ IV \leq 6$	$\%DE \geq 70$
Jelly	$IV \geq 6$	$58 \leq\ \%DE \leq 65$

Table 6.2: Results of deterministic optimization for three different types of fruits

	Product type	T (℃)	pH	Batch time (min)	$C_{pectin}(t_f)$ (g/l)	$IV(t_f)$ (dl/g)	$\%DE(t_f)$
Lemon							
	Base	80	2.37	259	10.86	6.68	71.7
	Jam	84	2.87	392	10.41	5.78	71
	Jelly	75	1.61	320	10.03	6.08	64.4
Orange							
	Base	82	2.51	261	12.05	4.90	66.1
Lime							
	Base	80	2.42	250	12.13	6.51	71
	Jam	85	2.85	353	11.84	5.73	70.3
	Jelly	73	1.66	325	11.14	6.06	64.7

The optimal values of temperature and pH are consistent with the results from the experiment-based optimization using response surface methodology (Gan and Latiff 2011; Masmoudi et al. 2008). Note that, the optimized batch time can be changed by the scale of the experiment as well as other design variables. However, the final concentration and critical quality attributes vary with the type of fruit. Fruits with higher total pectin content (orange or lime) result in a higher extraction yields. The same is seen for IV and $\%DE$, when it is extracted from fruits with higher initial values (lemon or lime) higher final values of these variables are observed.

In the cases of jam and jelly, the deterministic optimization is performed for lemon and lime only since orange has the lowest initial values of IV and $\%DE$, which are about the same as their lower bounds. The results from the deterministic optimization for lemon and lime peels show the same general trends of operating variables and outputs, for both jam and jelly cases. Hence, only lime peel is considered for the illustration of the optimization results. Figure 6.1 and Figure 6.2 show the profiles of outputs resulting from the deterministic optimization of the jam and jelly cases.

In the jam case, the optimized values of temperature and extraction time increased compared to the baseline scenario (see Table 6.2). This is necessary to induce the large reduction in IV, needed to satisfy the final IV constraint, as shown in Figure 6.1A. Moreover, when comparing with the base case, the optimized value of pH increases to compensate for the effects of increased temperature and extraction time on $\%DE$. This leads to meeting the lower bound of final $\%DE$ corresponding to 70% (Figure 6.1B). However, due to the excessive degradation by the high temperature and long extraction time, the final concentration is reduced by 2.4 % compared to that in the base case (Figure 6.1C). As shown in Table 6.2 and Figure 6.2B, in the jelly case, the final $\%DE$ is reduced below its upper limit through the large decrease in pH and increases in extraction time compared to the base case. Meanwhile, the optimized value of temperature becomes lower to compensate the effects of increased acidity and extraction time on IV in satisfying the lower limit of the final IV (Figure 6.2A).

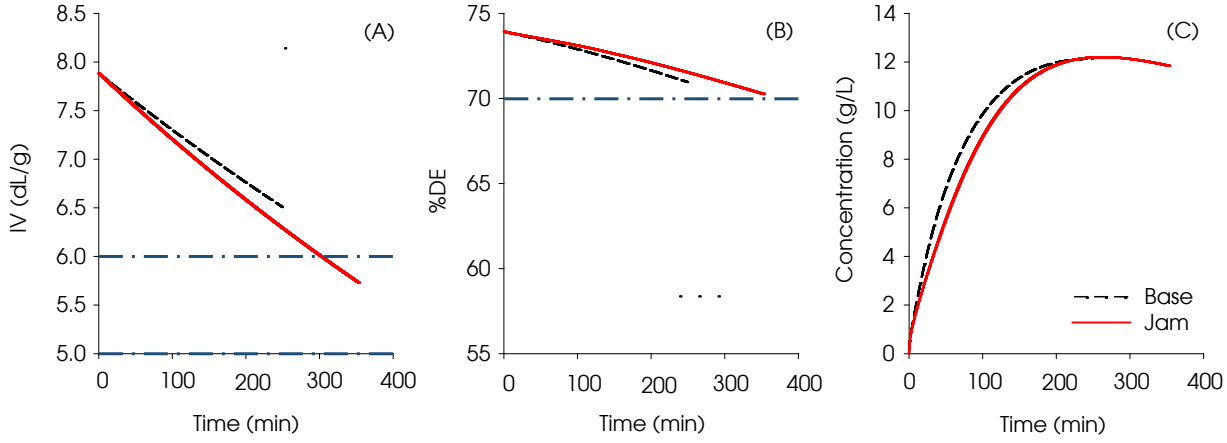

Figure 6.1: (A) IV, (B) $\%DE$ and (C) pectin concentration, resulting from the optimization of base case (dashed) and jam case (red) for a lime extraction. The required constraints of the final IV and $\%DE$ are shown (blue dotted).

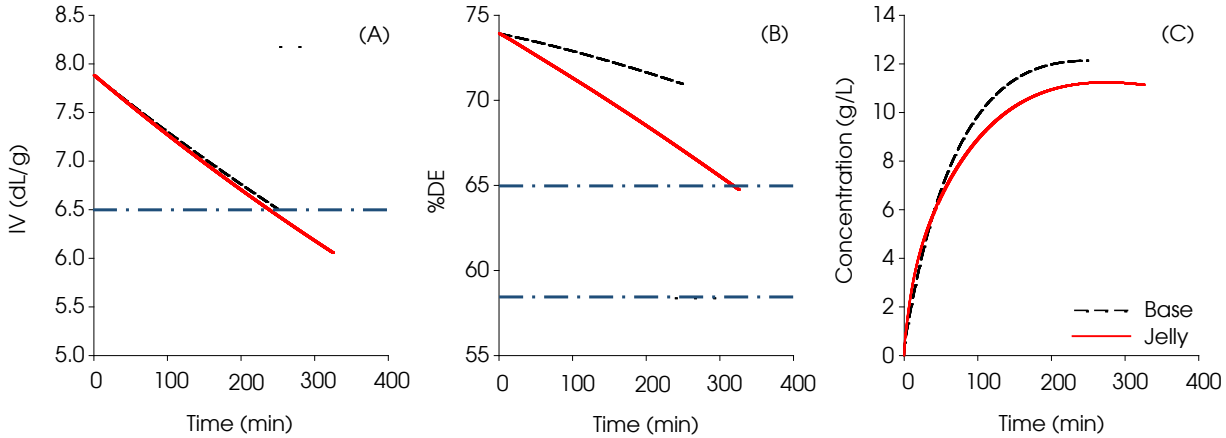

Figure 6.2: (A) IV, (B) $\%DE$ and (C) pectin concentration, resulting from the optimization of base case (dashed) and jelly case (red) for a lime extraction. The required constraints of the final IV and $\%DE$ are shown (blue dotted).

Since an increased acidity causes not only the hydrolysis but also the degradation to occur faster, the final pectin concentration in the jelly case is smaller (about 8.2 %) than that in the base case (Figure 6.2C).

As reported by Andersen et al. (2017), a higher IV and $\%DE$ can be obtained at a lower temperature and a higher pH. In addition, since the change in temperature mainly influences IV, the optimized value of temperature varies with the constraint of the final IV; in the jam (jelly) case, temperature significantly increases (decreases) to satisfy the upper (lower) bound on the final value of IV. On the other hand, pH has a strong effect on $\%DE$ but little effect on IV; in the jelly (jam) case, pH decreases (increases) significantly to satisfy the upper (lower) bound on the final value of $\%DE$. Therefore, the optimal operation condition is varied by not only the desired pectin quality depending on specific applications but also the initial condition of feedstock (Table 6.2). Orange peels are an abundant source for pectin (pectin load), and are more widely available than lime and lemon peels due to the different sizes of their respective juice industries. They production of pectin derived from orange is stil relevant, eventhough the operation of the extraction process requires the alleviation of lower bounds on IV and $\%DE$ due to the low initial conditions of them. Despite the low initial quality of orange, the pectin extracted from this material has an important role in increasing the production's product output, through standard/custom blending with higher quality peels, because it contains a high amount of total pectin content.

6.3 Robust optimization

In section 6.2, the optimal operating conditions are obtained in a deterministic manner using the mean values of initial conditions for the individual types of fruit. According to the model analyses in Chapter 5, the initial values of IV, $\%DE$, and pectin content in the peel are significant parameters for both the yield and the product quality. However, there are significant uncertainties in the initial conditions for even the same type of feedstock (see Table 5.3) leading to variances in the outputs and process performance as represented in the Monte-Carlo simulation results in Chapter 5. Especially, the final pectin characteristics are highly affected by their initial condition, and with the optimal operation condition obtained from the deterministic optimization, the inequality constraints that represent the desired quality of extracted pectin can be violated due to the uncertainty. However, an assessment of new raw material before every extraction process may be expensive and impractical. In this regard, the operation optimization should take into account the inherent uncertainty of feedstock.

The robust optimization for the pectin extraction process is examined to handle the uncertainty in a worst-case sense. The performance of the resulting robust operation condition is compared with that of the optimal operation condition obtained from the deterministic optimization through Monte Carlo simulations. As in the deterministic optimization, the worst-case scenario optimization is conducted for the

Table 6.3: Result of robust optimization of pectin extraction from the lime peel for the jam case (final $5 \leq IV \leq 6$ and final $\%DE \geq 70$) and simulation results for the worst cases.

Optimization	T (℃)	pH	Batch time (min)	Perturbation in initial conditions	$C_{pectin}(t_f)$ (g/l)	$IV(t_f)$ (dl/g)	$\%DE(t_f)$
Deterministic	85	2.85	353	$\mu + \sigma$	14.31	6.09	71.6
				$\mu - \sigma$	9.37	5.54	69.7
Robust	87.5	3.01	334	$\mu + \sigma$	13.85	6	72
				$\mu - \sigma$	9	5.45	70.2

Table 6.4: Result of robust optimization of pectin extraction from the lime peel for the jelly case (final $IV \geq 6$ and final $58 \leq \%DE \leq 65$) and simulation results for the worst cases.

Optimization	T (℃)	pH	Batch time (min)	Perturbation in initial conditions	$C_{pectin}(t_f)$ (g/l)	$IV(t_f)$ (dl/g)	$\%DE(t_f)$
Deterministic	73	1.66	325	$\mu + \sigma$	13.45	6.37	65.7
				$\mu - \sigma$	8.77	5.8	64
Robust	66	1.48	326	$\mu + \sigma$	12.43	6.61	64.9
				$\mu - \sigma$	8.12	6.02	63.2

pectin extraction from the lime. For the worst-case scenarios, two initial conditions are chosen as plus and minus a standard deviation from their mean values. In the robust optimization, the operating variables are optimized to satisfy given inequality constraints and maximize final pectin concentrations for the worst of the cases.

Table 6.3 and Table 6.4 summarize the robust optimization results together with the results obtained by simulating the process with the optimal process conditions obtained through deterministic optimization.

For both jam and jelly cases, the changes in the optimized values of temperature and pH by the robust optimization show the same general trends with those of the deterministic optimization and the optimized batch time from the robust optimization remains close to those from the deterministic optimization (Table 6.2). In the robust optimization for the jam case, the optimized value of temperature rises to reduce IV from the higher initial value ($\mu + \sigma$) to its upper bound. Besides, when the initial conditions are set as the lower bound ($\mu - \sigma$), the decrease of lower initial $\%DE$ is impeded by the increase in pH due to the lower bound for final $\%DE$. In the jelly case, the robust optimization gives lower values of optimal temperature and pH to meet the desired quality for the worst cases as shown in Table 6.4. As the initial value of $\%DE$ increases ($\mu + \sigma$), stronger acidity is required to induce faster de-esterification and satisfy its upper bound. On the other hand, when the initial value of IV decreases ($\mu - \sigma$), the final value of IV can be maintained above its lower bound as a result of the decreased effect of temperature on the reduction in IV. Note that these constraints are violated for the worst case simulations using the optimal operating variables obtained from the deterministic optimization as presented in Table 6.3 and Table 6.4. Therefore, through robust optimization, the adverse effect of uncertainties in initial conditions (i.e., constraint violation) can be reduced but, it involves a decrease in the objective function value (i.e., final pectin concentration).

Monte Carlo simulations are carried out, with the information of the lime peel (Table 5.3), to test the performances of the two optimal operating conditions resulting from the deterministic and robust optimizations. Just as performed before in Chapter 5, the u_{peel} variables are assumed to follow normal distributions, and the calculated mean and covariance matrix (for each fruit) are used to represent the input uncertainty space (a multivariate normal distribution) used for the random sampling. As a result of the Monte Carlo simulations with 10000 random samples, 58 % of all cases manifest the violation of the constraints for the production of jelly, while using the deterministic optimal operating conditions. When adopting the conditions from the robust optimizations, 24 % of all cases violate the required constraints. Using the robust operating conditions, the possibility of undesired constraint violation decreases, however, the averaged final pectin concentration becomes lower (by about 7.6 %). In the more extreme cases, such as $\mu \pm 2\sigma$ or 3σ, the robust optimization can result in more robust and aggressive solutions to satisfy the requirements of the extracted pectin at a higher expense of the yield. Besides, it may not be possible to find a robustly feasible solution strictly satisfying all constraints in extreme cases. Since there is a trade-off between the maximization of the objective function value and the satisfaction of the constraint, adequate worst case scenarios need to be iden-

tified in terms of cost for the raw material and downstream processing, and product price (Gorissen et al. 2015).

6.4 Concluding remarks

In this chapter it was shown how the conventional batch extraction process of pectin from citrus peels can be optimized, considering not only the variability of the raw material but also meeting the criteria for the quality constraints set for the desired pectin product. The application illustrates the potential use of model-based robust optimization for another process in which raw material variability can be an issue to achieve the sought product quality. The deterministic and robust optimizations are performed to determine optimal operating conditions producing pectin with the desired quality for applications in jam and jelly, from three different types of fruits. This provides the operational strategy the systematic basis for determining optimized process conditions and a tool to help with the selection of raw material and production planning.

CHAPTER 7

State Estimation

This chapter addresses the algorithms that combine a process model and physical measurements, in order to obtain a corrected estimate of the desired critical quality attributes. Section 7.1 describes the methods and concepts necessary to understand how this combination between model and measurements can be made. In Section 7.2 the expected difficulties of applying the model in different situations are exposed to highlight the types of problems these mathematical constructs can help correct. Section 7.3 is dedicated to the implementation of the extended Kalman filter method to actual full-scale data in the context that is relevant to the operational strategy. Issues such as correcting poor initial conditions and dealing with the differences in parameters of the pectin extraction process are addressed.

7.1 Concepts and methods

For complex systems such as bio-processes, the availability of measurement devices and techniques that characterize the process key performance indicators is limited. This can be either due to technical impossibility or due to unfeasible economical and

logistic arguments to measure these variables directly. Moreover, analytical technologies in the bio-based field measure variables with more uncertainty than those in other industries. For these reasons, the use of soft sensors has become a promising field in bio-based production (Sagmeister et al. 2013b).

These soft sensors exploit the mathematical relationship between the output and the unmeasured (internal) variables, such that the latter is reconstructed from measurements related to the former. These could be expressed either through empirical or mechanistic models (Luttmann et al. 2012). The terms *observers* and *state estimators* are very often used interchangeably throughout the literature (Mohd Ali et al. 2015). However, while the use of *observer* is commonly applied to deterministic cases, the *state estimators* are usually employed in a stochastic application, when noise (process and/or measurement) are incorporated in the algorithm.

The state estimation problem requires knowledge of the process complexity and model limitations. This implies that an appropriate selection of states to be estimated, for the available measurements, has to be assessed. The necessary elements of the state estimation routines are the existence of a model representation of the different states (x), which takes into account the known inputs (u), and the measured outputs (y) to correct any discrepancies between the measured outputs of the process and those predicted by the model. The notation $\hat{x}$ is used to express the state estimate made by the observer, while x represents the real state of the (unmeasured) system. When the influence of noise and uncertainty is considered, this exercise can be seen as a probabilistic inference, where we attempt to estimate state variables (which are not necessarily known) using a set of observations in an optimal way.

Different approaches have been proposed throughout the years; notably Luenberger (1971) proposed a (homonymous) deterministic observer for linear systems. In stochastic linear problems, the optimal solution is provided by the Kalman filter (R. Kalman 1960). For non-linear systems, plenty of different approaches have been explored recently (Randek and C. F. Mandenius 2017), however the focus in this chapter is set on the *workhorse* of the industry and arguably the most ubiquitous sub-optimal estimator: the extended Kalman filter algorithm, for the ease of understanding the methods logic and its adequacy to most biochemical problems (Bogaerts and Wouwer 2004; Jørgensen et al. 2007).

7.1.1 State-space representation

A possible description of what the *state* of a dynamic system represents is: the required set of information that is sufficient to predict the behaviour of the system, when information of future inputs is available. The state-space consists of all the values that the system may assume. The system can be represented via a state vector (x), which is the collection of the required (information) variables. A state-space representation of a system structures the system model explicitly in respect to its external inputs (u) and outputs (y). The evolution of the system can be described via the time derivative of the state vector and the vector of the measured outputs as

shown in equation (7.1).

$$\begin{aligned} \frac{dx}{dt} &= f(x(t), u(t)) \\ y(t) &= h(x(t), u(t)), \quad x(0) = x_0 \end{aligned} \tag{7.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \longmapsto \mathbb{R}^n$ and $h : \mathbb{R}^n \times \mathbb{R}^m \longmapsto \mathbb{R}^p$ are the state derivative and output mappings assigned to each $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ vector pair. These mappings are smooth, meaning their partial derivatives with respect to the state variables $(x_1, ..., x_n)$ exist for any order and are continuous.

The representation in equation (7.1) is generic and can represent a non-linear case. For a linear system the state representation is composed of the same two equations and can be represented in matrix form in (7.2)

$$\begin{aligned} \frac{dx}{dt} &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t), \quad x(0) = x_0 \end{aligned} \tag{7.2}$$

where matrix A is called the dynamics (or system) matrix, B is called the input (or control) matrix, C is called the output (or sensor) matrix and D is called the direct term or (feedthrough) matrix. It is common that systems do not have a direct term, which implies that the input signal has no direct influence in the output (Aström and Murray 2009).

In bio-based processes it is very frequent the case that where we have non-linear input/control-affine systems and the equations in (7.1) can be reduced (Hangos et al. 2004). The definition for these systems is given below.

Definition 7.1.1 (Non-linear input-affine systems)

$$\begin{aligned} \frac{dx}{dt} &= f(x(t)) + \sum_{i=1}^{m} g_i(x(t))u_i(t) \\ y(t) &= h(x(t)), \qquad\qquad\qquad x(0) = x_0 \end{aligned} \tag{7.3}$$

where the $x \in M$, $u \in \mathbb{R}^m$ and $y \in \mathbb{R}^p$ are the state, input and output vectors with $f, g_1, ..., g_m$ as smooth nonlinear mappings defined in M. M is the state space and an open subset of $\mathbb{R}^n$.

7.1.2 Observability

For the development of a state observer, it is important to assess if the desired states can be estimated from the process model, system inputs, and physical measurements. The concept of observability is then a necessary, albeit not sufficient, condition for the success of an estimator implementation.

Observability can be intuitively understood as a system's property which permits the possibility of the inference of the state variables with the available outputs, for given system inputs in a finite time (Dewasme et al. 2013).

For the following linear system of equations

$$\begin{aligned} \frac{dx}{dt} &= Ax(t) + Bu(t) \\ y(t) &= Cx(t), \qquad\qquad x(0) = x_0 \end{aligned} \tag{7.4}$$

where matrix $A \in \mathbb{R}^n \times \mathbb{R}^n$, $B \in \mathbb{R}^m \times \mathbb{R}^n$ and $C \in \mathbb{R}^p \times \mathbb{R}^n$. If C is invertible it is possible to reconstruct the states x from the output measurements in a straightforward fashion. Given that we are not in that situation, the information in further time-derivatives of the outputs is assessed to verify if we have an observable system.

The system in (7.4) has the following observability matrix, as introduced in R. E. Kalman 1960.

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{7.5}$$

The time-derivatives up to $n-1$ are sufficient to qualify if enough information is gathered for the estimation, due to the Cayley–Hamilton theorem (Aström and Murray 2009).

Definition 7.1.2 (Linear (time-invariant) observable system) *A system is observable if and only if rank(O) the rank of the observability matrix is n*

$$rank(\mathcal{O}) = n$$

where n is the state vector dimension

This statement provides a mathematical condition of whether a set-up with a given model and measurement structure choice is observable.

7.1.2.1 Non-linear systems

The extension of observability to non-linear systems is not straightforward. While considering a system such as the one in (7.3) we can define one type of observability around the concept of distinguishability, according to Hermann and Krener (1977).

Definition 7.1.3 ($\mathcal{U}$-Indistinguishability) *Two different sets of points $x(t_0) = x1$ and $x(t_0) = x2$ in an open set $\mathcal{U}$ are indistinguishable if their system solutions $x(t)$, for all admissible defined $u(t)$, are identical.*

Definition 7.1.4 (Local Observability) *A system is said to be locally observable for a specific x_0 if said initial condition vector has an open neighborhood $\mathcal{U}$ that for every open neighborhood $\mathcal{V}$ of x_0 contained in $\mathcal{U}$ there are no indistinguishable points from x_0.*

Essentially, a system is found to be observable if this condition is satisfied at every point in the state space and for any valid input. Whilst this local observability concept defines that a system is observable if for all states exists at least one input that allows the discrimination of this state with all nearby states, by measuring the output, at a specific initial state and around its neighborhood.

This type of observability is very useful as it allows a practical evaluation. However, this requires the introduction to the differential geometry concepts of Lie derivatives:

$$L_f h(x) = \frac{\partial h(x)}{\partial x} f(x) \tag{7.6}$$

which for $i = 1, ..., n-1$ has the following property:

$$L_f^i h(x) = L_f(L_f^{i-1} h(x)) = \frac{\partial L_f^{i-1}}{\partial x} f(x) \tag{7.7}$$

For the systems considered, $L_f h(x)$ is the equivalent of the first-order time derivative of h. The local observability of a system can be assessed through analysing the singular values of the following observability matrix:

$$\mathcal{O}(x)_{p\cdot n \times n} = \begin{bmatrix} \frac{\partial}{\partial x_1} h(x) & \dots & \frac{\partial}{\partial x_n} h(x) \\ \frac{\partial}{\partial x_1}(L_f h(x)) & \dots & \frac{\partial}{\partial x_n}(L_f h(x)) \\ \frac{\partial}{\partial x_1}(L_f^2 h(x)) & \dots & \vdots \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1}(L_f^{n-1} h(x)) & \dots & \frac{\partial}{\partial x_n}(L_f^{n-1} h(x)) \end{bmatrix}, \qquad h(x) \in \mathbb{R}^p \tag{7.8}$$

This observability matrix establishes the link between the state vector x and the time derivatives of the measurements/outputs. This matrix is dependent of the number of states (n) and measurements (p) of the system. A practical way to assess if the system is observable is by the following theorem

Theorem 7.1.1 (Observability) *If the system satisfies the rank condition for* $\mathcal{O}(x_0) = n$ *then system* (7.1) *is locally observable at* x_0

Furthermore, the analysis provided by the non-linear observability methods can be extended to structural identifiability. Model parameters can be augmented to the state vector as constant state variables ($\frac{dx}{dt} = 0$), then it is possible to determine the identifiability through the rank check of the non-linear model augmented observability matrix (Villaverde and Banga 2017).

7.1.3 Kalman filtering

The general case for the linear discrete form of the Kalman filter is given to introduce the concepts of this state estimator. More profound and detailed information is

available in multiple books and other literature resources (e.g., Grewal and Andrews 2008).

Consider the following linear system

$$\begin{aligned} x_k &= Ax_{k-1} + Bu_k + w_k \\ y_k &= Cx_k + v_k \end{aligned} \tag{7.9}$$

where matrix $A \in \mathbb{R}^n \times \mathbb{R}^n$, $B \in \mathbb{R}^m \times \mathbb{R}^n$ and $C \in \mathbb{R}^p \times \mathbb{R}^n$. These transition matrices are here assumed as constant. Both w_k and v_k are mutually independent sequences of zero mean white Gaussian noise with joint covariance, for process and measurement, respectively. These noise variables vectors have the following covariance matrices

$$\begin{aligned} E[w_k w_k^T] &= P_k^w = Q \\ E[v_k v_k^T] &= P_k^v = R \\ E[w_k v_k^T] &= 0 \end{aligned} \tag{7.10}$$

Considering the initial conditions mean and covariance matrix

$$\begin{aligned} E[x_0] &= \hat{x}_0 \\ E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T] &= P_0^x \end{aligned} \tag{7.11}$$

the filtered estimated is obtained through the minimization of the mean-square error between the state and the estimate. This is represented in Figure 7.1, where $\hat{x}$ is the state estimate. The algorithm is typically formulated in two steps, prediction and update. An *a priori* estimate and its covariance is computed in the prediction step with the information up until that time:

$$\begin{aligned} \hat{x}_{k|k-1} &= A\hat{x}_{k-1|k-1} + Bu_k \\ P_{k|k-1}^x &= AP_{k-1|k-1}^x A^T + Q \end{aligned} \tag{7.12}$$

as soon as a new measurement is available, at time k, this values can be updated in the update step, which requires the computation of the Kalman gain K_k

$$\begin{aligned} K_k &= P_{k|k-1}^x C^T (CP_{k|k-1}^x C^T + R)^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}) \\ P_{k|k}^x &= (I - K_k C)P_{k|k-1}^x \end{aligned} \tag{7.13}$$

This algorithm does not represent a big expenditure of data storage or computational power as it merely requires calculations at each time point k, demanding only information from the immediate previous time point $k-1$ and k itself.

7.1.3.1 Continuous-discrete extended Kalman filter

In nonlinear processes, the most common approximate solution to the optimal state estimation problem is the extended Kalman filter (EKF). A particular sub-optimal

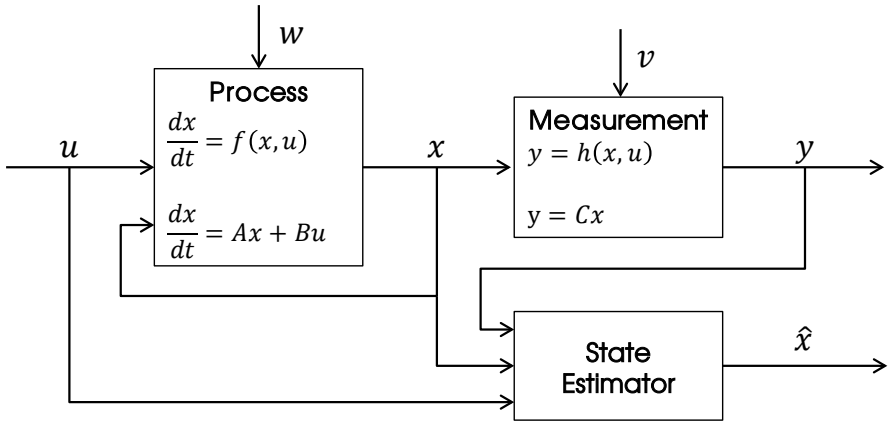

Figure 7.1: Diagram of a simplified block representation of the observer concept with state space notation for both the linear and non-linear case

form of the Kalman filter is the continuous-discrete (CD-EKF) that is applied in systems in the (7.14) form, where the system is continuous non-linear time-invariant with measurements in discrete time.

$$\begin{aligned} \frac{dx}{dt} &= f(x(t), u(t)) + w(t) \\ y(k) &= h(x(k)) + v(k) \end{aligned} \tag{7.14}$$

where the $x \in M$, M is the state space and an open subset of $\mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, $w \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ are the state, input, output, the process and measurement noise variables vectors, respectively.

The noises are assumed Gaussian random with zero mean $E[w(t)] = E[v(k)] = 0$ and are uncorrelated $E[w(t)v^T(k)] = 0$.

The EKF sub-optimal solution to estimation comprises the same two steps (prediction and update) as the linear Kalman filter, however it provides an approximation of the non-linearities of the system (7.14) through linearization, typically, around its last state estimate. The successful implementation of the solution relies on the validity of linearizing, as an adequate representation of the system, in the entire state domain. The CD-EKF can diverge if approximating with sequential linearization is not sufficiently accurate.

The continuous-discrete variant entails the prediction step to be made with the use of a continuous model, while the update step is calculated discretely at each time-point where a measurement is available. For the prediction step, the state propagation requires the integration of the differential equation $k-1$ to k

$$\frac{d\hat{x}}{dt} = f(\hat{x}(t), u(t)) \tag{7.15}$$

whilst the covariance prediction requires the computation of the Jacobian A matrix at every integration time step of the ODE solver.

$$A(t) = \frac{\partial f}{\partial x}\Big|_{\hat{x}_{k|k-1}(t),u_{k-1}} \tag{7.16}$$

and to predict the covariance ($P^x_{k|k-1}$), in continuous fashion, it is made through its Riccati equation form

$$\frac{dP}{dt} = A(t)P + PA(t)^T + Q \tag{7.17}$$

with initial condition as the previours *posterior*

$$P_0 = P^x_{k-1|k-1} \tag{7.18}$$

The update equations are equal to the ones in (7.13) with the particularity of defining matrix C using the Jacobian of the measurement function.

$$C_k = \frac{\partial h}{\partial x}\Big|_{\hat{x}_{k|k-1}(t),u_k} \tag{7.19}$$

$$\begin{aligned} K_k &= P^x_{k|k-1}C^T(CP^x_{k|k-1}C^T + R)^{-1} \\ \hat{x}_{k|k} &= \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}) \\ P^x_{k|k} &= (I - K_kC)P^x_{k|k-1} \end{aligned}$$

The tuning of the parameters was performed empirically on a trial and error basis for the simulation case in the next section. However, for the implementation with real data the tuning and initialization of the filter is made following Schneider and Georgakis (2013) rationale for a more systematic design. Particular focus is given to the choice of the design parameters initial state P^x_0 and process noise Q covariance matrices. The measurement noise covariance matrix R is considered available from the measurement system and can be provided by the analytical experts, leading to the following diagonal matrix:

$$R = k_R \cdot diag(\sigma_p^2) \tag{7.20}$$

where σ_p is the standard deviation of measurement p and k_R is a corrective factor diagonal matrix which can be adjusted if the measurement uncertainty is expected to be higher than the considered σ_p. In this assessment, we assume the values for uncertainty provided are adequate and time-invariant ($k_R = 1$)

Regarding the initial state guess ($\hat{x}_0$) and its associated uncertainty (P^x_0), their exact values are often difficult to known in practice. However, if the states are related to critical material attributes that can be assessed (as seen in Chapter 4), then an initial estimate of the states and an uncertainty related to it is available and depends on the quality assessment approach is selected.

$$P^x_0 = k_{P^x_0} \cdot diag(\sigma^2_{x_0}) \tag{7.21}$$

if P_0^x is deemed to be too optimistic for the actual initial state guess discrepancy, then the corrective factor vector $k_{P_0^x}$ may be used to adjust the uncertainty around any of the initial state guesses.

As for the process noise covariance (Q), this uncertainty is related to both external disturbances to the process and internal errors associated to deficient knowledge of the system or errors derived from numerical approximations. Even though random process noise exists, the parametric uncertainty associated with the model identification is assumed to be what plays a dominant role, even when the model structure is adequate to follow the real system. From the results of parameter estimation (as seen in Chapter 5) the uncertainty of the estimates for each parameter is found. Thus, according to the Schneider and Georgakis (2013) approach, one can calculate Q from the covariance of these parameters if consider Equation (7.14) to include the stochastic part of the estimated parameter.

$$\frac{dx}{dt} = f(x(t), u(t), \hat{p}) + J_{\hat{p}}(t)\delta p \tag{7.22}$$

where δp represents the stochastic part of the estimated parameter with covariance $C_{\hat{p}}$ and $J_{\hat{p}}(t)$ is the linear term of the Taylor expansions calculated as

$$J_{\hat{p}}(t) = \left(\frac{\partial f}{\partial p}\right)_{x(t),u(t),\hat{p}} \tag{7.23}$$

this leads to the equality between the process noise variable and the uncertainty term in Equation (7.22), which then allow us to calculate the process noise covariance matrix (Equation (7.10))

$$\begin{aligned} w(t) &= J_{\hat{p}}(t)\delta p \\ E[w(t)w(t)^T] &= (J_{\hat{p}}(t)\delta p)(J_{\hat{p}}(t)\delta p)^T = J_{\hat{p}}(t)C_{\hat{p}}J_{\hat{p}}(t)^T \\ Q(t) &= k_Q \cdot J_{\hat{p}}(t)C_{\hat{p}}J_{\hat{p}}(t)^T \end{aligned} \tag{7.24}$$

The matrix Q is often used as time-invariant, but it can be seen that it could be considered to be time-varying if $J_{\hat{p}}$ is time-varying, even if the covariance $C_{\hat{p}}$ is time-invariant. Moreover, this matrix is considered a "tunable" parameter whilst the R matrix is typically maintained at the measuring systems specifications. Hence by manipulating the factor k_Q *ad hoc* to achieve the desired performance of the state observer is common practice (Schneider and Georgakis 2013).

7.2 Discrepancies between model simulation and applications

In the previous simulation scenario, it was assumed that no mismatch exists between the plant dynamics and the observer model. However, it can be seen from Figure 7.2 and Figure 7.3 that it is not always the case in actual extractions.

Figure 7.2 shows the linear fit within fruit groups and the 95% confidence interval for the considered laboratory data, previously described in section 4.3.1. This highlights how the kinetics of the evolution of the critical quality attributes IV and $\%DE$ through time might differ from peel-to-peel, or at the very least, between subgroups of peels. This is evident in the IV case, where it can be seen that the degradation of this variable is slower than lemon and orange. For $\%DE$ the dynamics have more similarity across the different fruits.

The results in Figure 7.3 are obtained by using as initial condition the pectin content obtained experimentally for these peels in a laboratory scale extraction $C^0_{pectin} = 86.7$ kg/m^3 and $C^0_{pectin} = 38$ kg/m^3 for the lemon and orange peels, respectively. It is seen in that the simulations, which adopt parameters identified with a lime peel, have a mismatch from the actual data in the lemon case. The actual data suggests a slower diffusion than the ones estimated for the lime peel by Andersen et al. (2017), implying also bigger influence of the temperature in the activation energy of this phenomena. However, for the orange case the simulation in close the experimental data, revealing a close performance of this peel to the lime used for parameter estimation.

The discrepancies may also occur due to the equipment/scale induced bias. A Monte-Carlo simulation, performed similarly as described in section 5.1.2.4, with vector input uncertainty consisting of the u_{peel} variables, with lime statistics (see Table 5.2), together with the uncertainty regarding the operational conditions of all the considered fullscale extractions, here the mean and standard deviations of the mass of peel, volume of extraction, temperature and pH were considered for the sampling procedure. Figure 7.4 A) illustrate the differences that arise from model simulation and data at full-scale. The model, as it was identified, does not accurately portray the extraction of pectin at fullscale. It under-estimates the amount of pectin

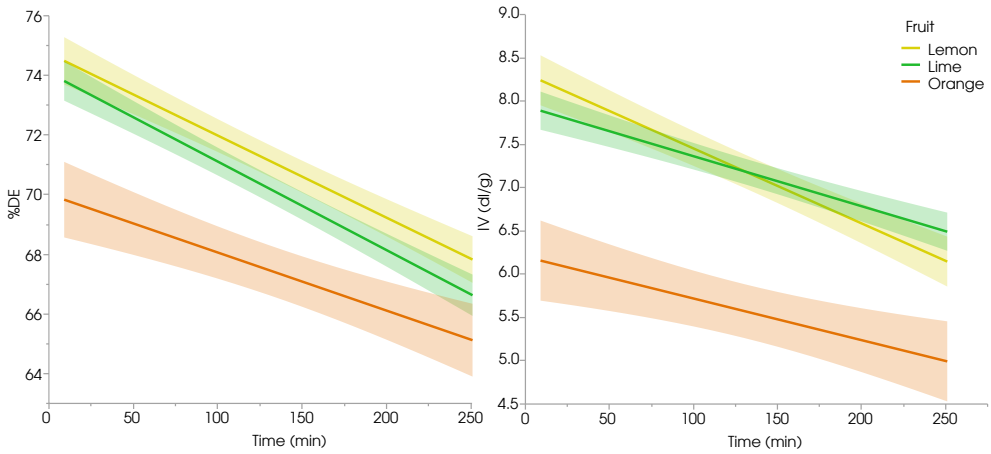

Figure 7.2: $\%DE$ and IV evolution in labscale extractions. Colored and fitted by type of fruit.

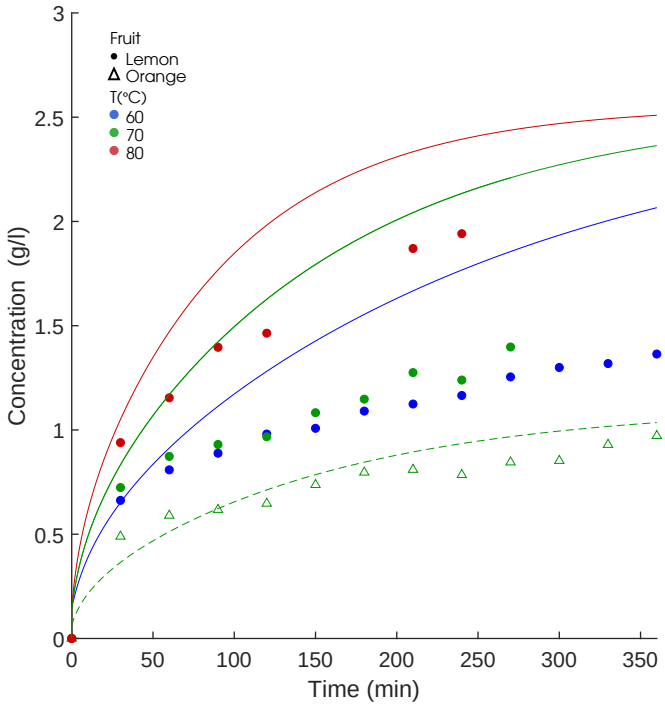

Figure 7.3: Simulation of Andersen et al. (2017) model, with re-estimated parameters, and data from T=70-80°C water experimental runs for lemon and orange peels.

that is extracted, which points out to the possibility of the over-representation of the phenomena of concentration degradation. When this term is disregarded from the model, the prediction improves greatly. This is represented in Figure 7.4 B). It is also not observable from the data trends any indication that this phenomenon occurs at this fullscale, at these conditions, nor was it confirmed by the experts at CP Kelco. Nevertheless, there is still discrepancy when looking at the earlier stages of the extraction, where the data exhibits a faster washing stage than the one portrayed through the model.

The peel used in these extractions is from a lime fruit, similarly to the peel used in the parameter identification. However, there are some mismatches at IV and $\%DE$ profiles. Figure 7.5 A) exhibits a model simulation that has an appropriate initial condition, but a faster de-esterification rate than the real system. In contrast, Figure 7.5 B) reveals that the IV model has adequate kinetics but adopting the lime group mean value of IV_0 shows to be insufficient.

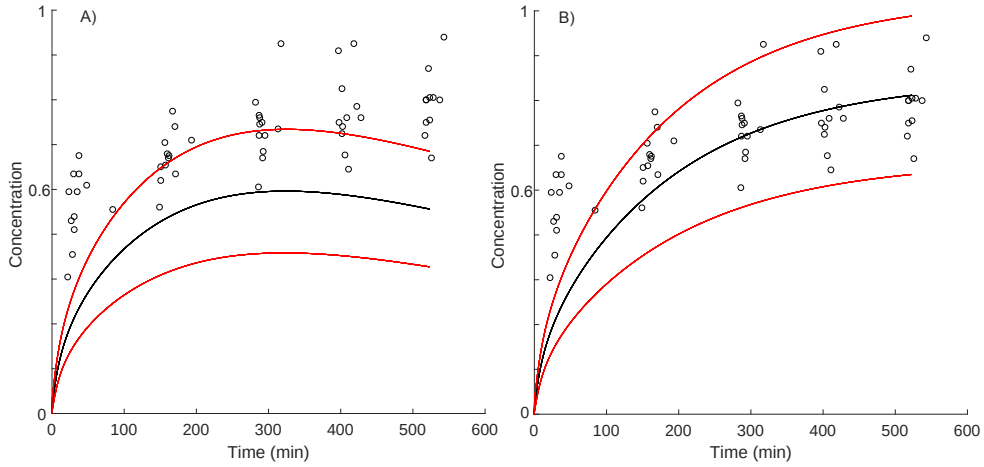

Figure 7.4: A) Monte-carlo simulation (with 95% confidence interval) of Andersen et al. (2017) model (with re-estimated parameters) and fullscale concentration data. B) without accounting the concentration degradation effect. The Y axis are scaled to the +95% confidence interval at t_f for Figure B).

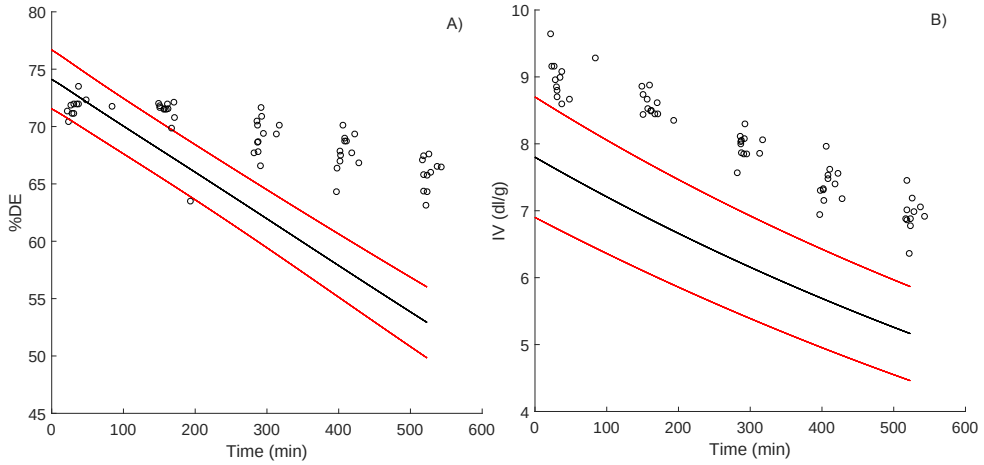

Figure 7.5: A) Monte-carlo simulation (with 95% confidence interval) of Andersen et al. (2017) model (with re-estimated parameters) and fullscale data of A) $\%DE$ B) IV.

7.3 Batch quality monitoring and forecast

In this subsection, the previously introduced continuous-discrete extended Kalman filter is assessed for its prediction qualities, when using data from production scale extraction batches. The CD-EKF is implemented so that the current estimate may be corrected from the available measurements, but more importantly so that it provides an improved predictor. In the latter, the states and estimator errors are propagated from the current measurement time (k) to the batch termination time (t_f), without further corrections. The predictive batch monitoring capability is an important part of the proposed operational methodology as it serves to early detect the deviating batch behaviour due to deviations from the initial conditions or discrepancies in the dynamics.

The extractions described in section 2.4.4, are used to showcase the development of the forecast tool. The data is used as such: six extractions are used to tune the Kalman filter and the other six are used as a test. Figure 7.6 illustrates the split between calibration and validation samples.

In this section, the CD-EKF algorithm is investigated in the context of an improving predictor role. The following subsection 7.3.1 analyses the systematic implementation of CD-EKF as means of correction of state estimate with measurement update, followed by propagation of estimates and error until t_f. However, this implementation raises some concerns which are addressed in the subsequent subsection 7.3.2.

The analysis starts from the Andersen et al. (2017) with the re-estimated parameters in Chapter 5 (see Table 5.11). However, the $\%DE$ variable is decoupled from the pectin concentration for observability purposes, since this variable is measured independently and the inference of both do not benefit from the measurements of each-other. Thus, equations (2.35) and (2.36) which are used to calculate $\%DE$ are simplified to the previously assessed first order reaction of the de-esterification.

$$\frac{d\%DE}{dt} = -k_{de-esterification} \cdot \%DE(t) \tag{7.25}$$

Throughout this section, the concentration results are scaled to the +95% confidence interval at t_f resulting from the Monte Carlo simulation (see Figure 7.4 B).

7.3.1 Improved predictor with state estimation

In this system, the discrete measurements of the key performance indicators are used for updating the forecast throughout the batch we have the following state and measurement vectors:

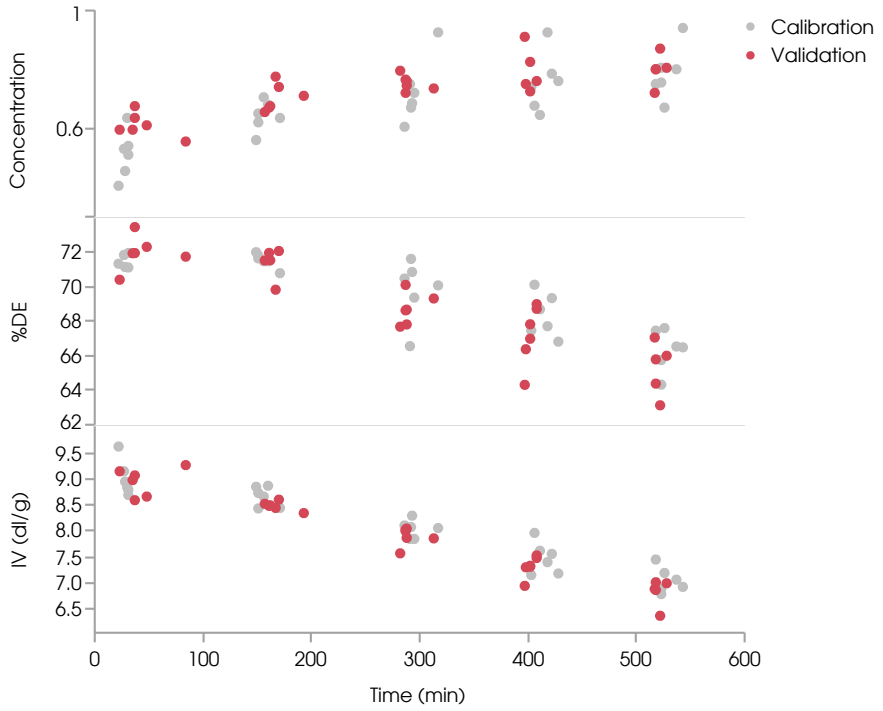

Figure 7.6: Fullscale batch extraction profile data for pectin concentration, $\%DE$ and IV. Colored by calibration or validation membership.

$$x = \begin{bmatrix} C^{peel}_{pectin,x=0} \\ \vdots \\ C^{peel}_{pectin,x=L} \\ C_{protopectin} \\ C^{bulk}_{pectin} \\ \%DE \\ IV \end{bmatrix}, \qquad y = \begin{bmatrix} C^{bulk}_{pectin} \\ \%DE \\ IV \end{bmatrix} \tag{7.26}$$

since the states $\%DE$ and IV are decoupled from the concentration, and their measurements are not used for any further state estimation, they function as detectors and their observability is trivial

$$\mathcal{O}(x_{IV})_{1\times 1} = \mathcal{O}(x_{\%DE})_{1\times 1} = 1 \tag{7.27}$$

For the concentration state, a discretization with $N = 10$ peel states yields state vector has $dim(x_C) = 12$ and $\mathcal{O}(x_C)_{1\cdot 12\times 12}$. To assess the observability in this set-up

we first start to build the Lie derivative vector

$$q_{x_C} = \begin{bmatrix} C_{pectin}^{bulk} \\ \vdots \\ \frac{d^{n-1}}{dt^{n-1}} C_{pectin}^{bulk} \end{bmatrix} \tag{7.28}$$

which leads to the following observability matrix

$$\mathcal{O}(x_C)_{12\times 12} = \begin{bmatrix} \frac{\partial}{\partial C_{pectin,x=0}^{peel}} C_{pectin}^{bulk} & \cdots & \frac{\partial}{\partial C_{pectin}^{bulk}} C_{pectin}^{bulk} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial C_{pectin,x=0}^{peel}} \frac{d^{n-1}}{dt^{n-1}} C_{pectin}^{bulk} & \cdots & \frac{\partial}{\partial C_{pectin}^{bulk}} \frac{d^{n-1}}{dt^{n-1}} C_{pectin}^{bulk} \end{bmatrix} \tag{7.29}$$

The rank of the observability matrix is calculated for the course of the simulation time, using as process conditions and peel inputs the same mean values used for the simulation of Figure 7.4. The normalized rank of the observability matrix can be seen in Figure 7.7. It shows that for the $N = 10$ model here considered and used by Andersen et al. (2017), the estimability structure (model, inputs and measurements) does not have full rank throughout the entire trajectory of the batch, since the condition $rank(\mathcal{O}(x_C))/n = 1$ is not satisfied. This implies that the invertibility of the observability matrix during the batch time is not ensured.

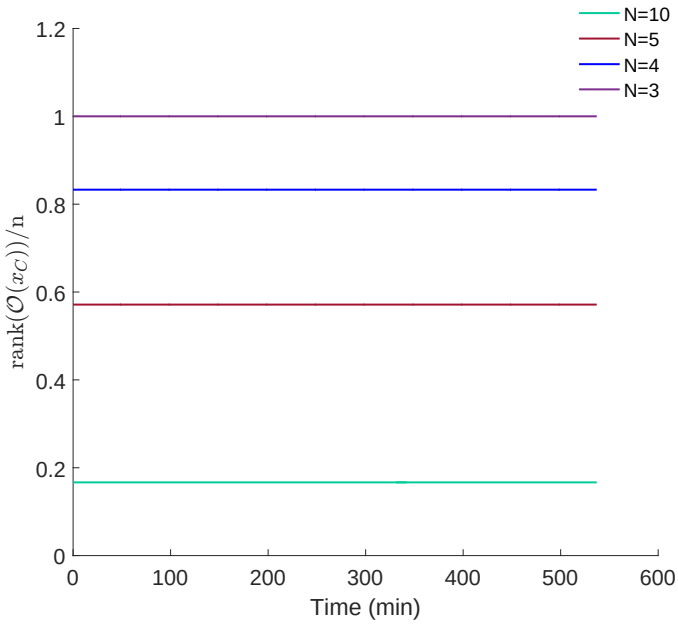

Figure 7.7: Normalized observability matrix rank throughout the course of the extraction simulation for the cases of discretization N=10, N=5, N=4 and N=3.

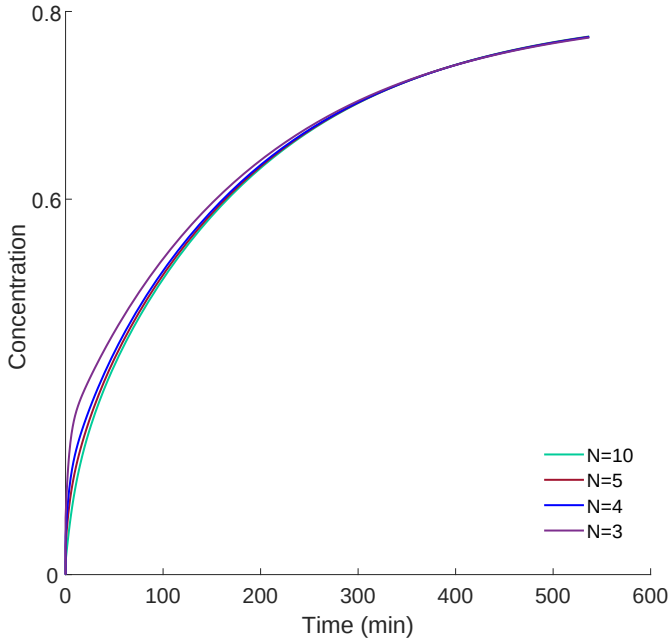

Figure 7.8: Pectin concentration in bulk simulation for the cases of discretization N=10, N=5, N=4 and N=3.

The constant value of $rank(\mathcal{O}(x_C))/n = 0.1667$ represents a rank of 2 for the during all the simulation period. The existence of N internal peels states hinders the observability. This is assessed by selecting different N and analysing the rank of the observability matrix for that discretization. For example, for $N = 5$ the normalized rank is 0.5714 (rank of 4 in 7 states), while for $N = 4$ the system has rank 5 in 6 states. For a minimum $N = 3$, which comprises a node in the center of the peel slab, a half-way node and the point at the interface we have full-rank. However, the simulations numerical accuracy suffers in this less discretized system as it can be seen in Figure 7.8. The filter is still attempted for $N = 10$, despite the non-observability of the internal peel nodes the application of a model based observer could still provide better estimates than simply applying the pure model. The case for the observable $N = 3$ discretization is also assessed.

Since the systems for the desired states are not coupled, we benefit from having three filter implementations for each system. Initializing and tuning the filter parameters can be difficult and a highly tentative process. However, here the measurement noise covariance matrices are defined from the standard deviations ($\sigma_{measured}$) associated with each measurement. These values can be seen in Table 7.1 provided by the analytical method experts at CP Kelco.

The states initial guess and error of estimation covariance matrices are defined

Table 7.1: Measurements noise represented in terms of standard deviation ($\sigma_{measured}$)

Measured Variable	$\sigma_{measured}$
$C_{pectin,bulk}$	0.6 g/l
IV	0.1 dl/g
$\%DE$	0.95 %

from the information gathered on raw material in Chapter 4. The u_{peel} variable values for lime peel and their associated uncertainty are used (see Table 5.3).

Moreover, rather than arbitrarily initializing the process noise covariance, and from there tentatively change the values in the diagonal, the time-variant $Q(t)$ based on parametric uncertainty is used for the different states.

Based on this initialization, extraction batch 1 is used to illustrate the differences in performance between discretizations $N = 3$ and $N = 10$. It can be seen in Figure 7.9 the variance propagation of the extracted pectin concentration prediction from each measurement update stages. It is visible that both exhibit the behaviour of a typical continuous system with discrete measurements where a new measurement tends to reduce P^x and the propagation step where the process noise is added increases the state estimate error covariance matrix. However, both cases exhibit non-convergence in estimate error. The probable causes for this non-convergence of the error covari-

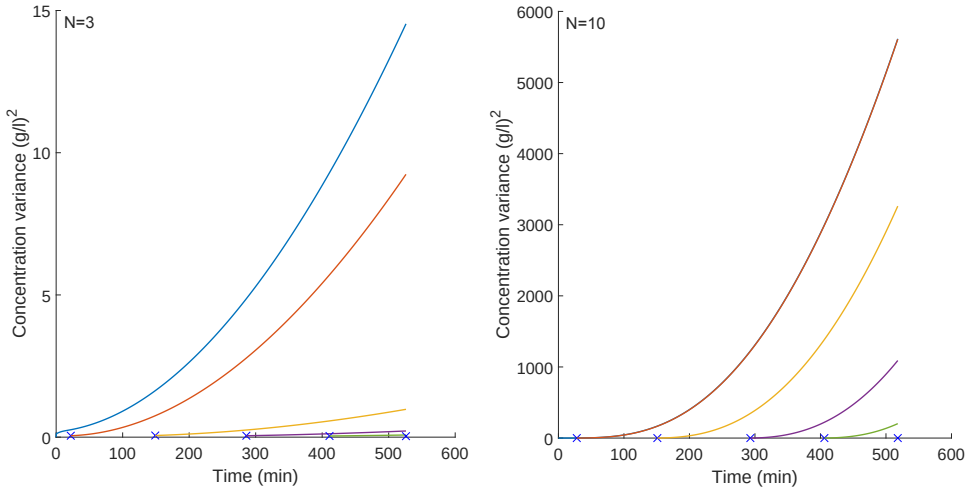

Figure 7.9: Propagation of the process estimate variance for the concentration of extracted pectin in the extraction 1 case. The value is extracted from the diagonal of the covariance matrix ($P^x_{k|k-1}$). The variance prediction is represented from each measurement update instance (×) until the end of the batch, for both cases of discretization N=3 and N=10.

ance can be related to the systems natural behavior or the poor observability with the given measurements and inputs.

There is a difference in the divergence rate between a system that meets the observability condition ($N = 3$) and the unobservable system ($N = 10$). In either case, the variance increases exponentially, however for the unobservable system this leads to an unacceptable value extremely quickly. The divergence for the observable system has to due to its poor conditioning to respect to inversion, which can be evaluated by the condition number:

$$cond\big(\mathcal{O}(x_C)_{5\times5}\big) = \frac{SV_{max}(\mathcal{O}(x_C)_{5\times5})}{SV_{min}(\mathcal{O}(x_C)_{5\times5})} = 2 \cdot 10^{13} \tag{7.30}$$

where $SV_{max}(\mathcal{O}(x_C)_{5\times5})$ and $SV_{min}(\mathcal{O}(x_C)_{5\times5})$ are the maximum and minimum singular values of the observability matrix $\mathcal{O}(x_C)_{5\times5}$, respectively. The condition number places emphasis on the smallest singular value. If a system is near singular, inversion of the observability matrix (or errors introduced by this inversion) is dominated by the smallest singular value of its observability matrix. Note that condition values closer to 1 represents a system with a better inversion of its observability matrix which leads to systems with superior state estimation. Considering the rule-of-thumb value for ill-conditioning is $\approx 2{\cdot}10^7$, this system is several orders of magnitude beyond that mark (Grewal and Andrews 2008; Lopez 1999).

Considering the case of $N = 3$ discretization, the performance of the filters with the above initialization can be seen for Figure 7.10 for the first measurement update iteration (k=1). The different initial simulations showcase the different *a priori* uncertainty for each variable, which is the joint result of initial u_{peel} expressed in the P_0^x and the added parametric uncertainty through the $Q(t)$ integration in equation (7.17). From the three filter models, concentration has the best initial prediction for an end-time estimate. However, the uncertainty regarding that estimate is quite substantial. The first measurement update worsens this estimate, because of the poor dynamic representation of the model in the fast diffusion (washing) stage.

The variables $\%DE$ and IV deal with different mismatches (as seen before in Figure 7.5). In the IV case, the filter model recovers from the poor initial condition estimate after $k = 1$ and the consequent model propagation provides a t_f estimate closer to the measured output. In the $\%DE$ filter model, it is not needed to correct for such bad initial guess. However, the lack of kinetic corrections means that the filter will not be able to converge quickly enough to the measured t_f output. In both these cases the calculated $Q(t)$ from the parametric uncertainty, puts a great deal of trust in the model. The estimates after the measurement update have a very small uncertainty attached to them as can be seen by the $\pm2\sigma_{prediction}$ lines. For a trustful predictor, the uncertainty related to these projections has to be adjusted such that even if the model is dynamically reliable, the inherent measurement uncertainty is captured with confidence. Thus, the filter model should be tuned to allow the overlap between measurements and estimates.

The tuning of the filter is done through optimization of the k_Q (from equation (7.24)) in the six calibration batches. The optimization is performed such that for the

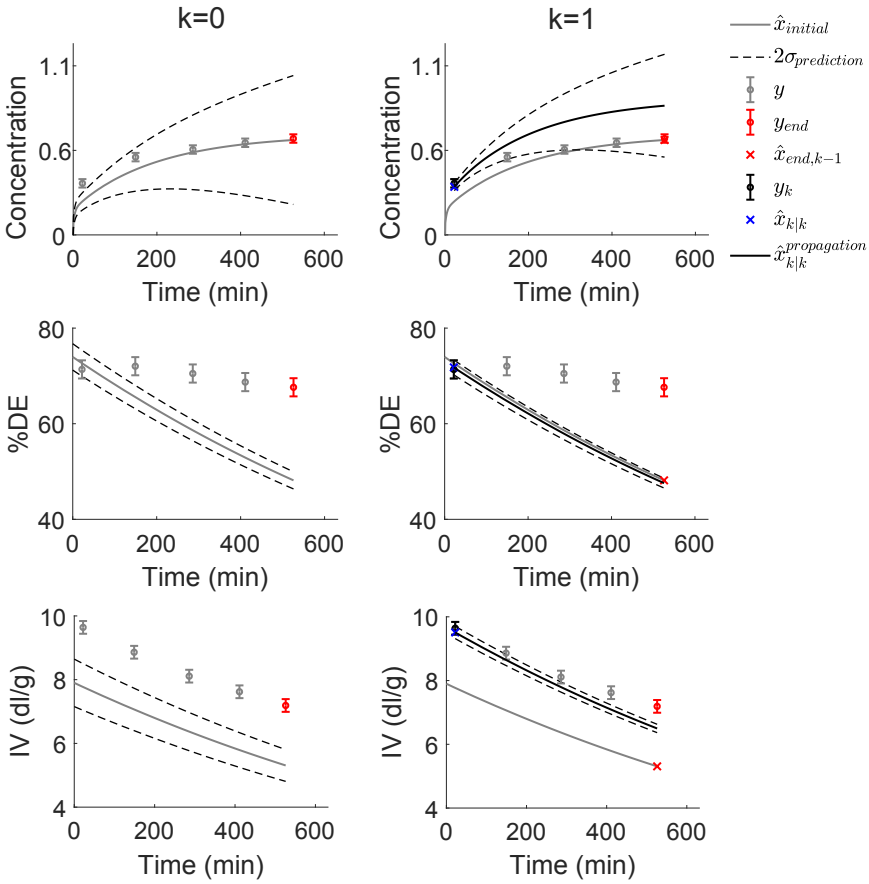

Figure 7.10: Predictor performance in extraction batch 6 for the different variables. The figures in the left column exhibit the initial prediction (k=0) whereas the right and after the first measurement update k=1. The measurements are represented with $2\sigma_{measured}$ error bar.

estimate at t_f, of the prediction after the first time-update (k=1), has an associated $\pm 2\sigma_{prediction}$ uncertainty of a maximum $\pm 3\sigma_{measured}$. This is deemed to be a good deviation for an end-batch estimate given an early sample for this implementation. The filters are run for the 6 calibrations batches and the Nelder-Mead simplex algorithm, implemented in MATLAB's *fminsearch* function, is used to find the minimum of the sum of squared errors as follows:

$$\begin{aligned} \underset{k_Q}{\text{minimize}} \quad & \sum_{n=1}^{6} \left(2\sigma^{t_f}_{n,k=1} - 3\sigma_{measured}\right)^2 \\ \text{subject to} \quad & k_Q \geq 1 \end{aligned} \tag{7.31}$$

where n is the extraction batch index. The value k_Q should however be higher than 1 such that the parametric uncertainty related to the previous parameter estimation is not undermined. The initial guess point of k_Q in the optimization is important, so an empirical examination can be made for one batch. The resulting factors are: $k_{Q,C} = 1$, $k_{Q,\%DE} = 5.12{\cdot}10^5$ and $k_{Q,IV} = 5.67{\cdot}10^4$. For the concentration filter, the uncertainty of the predicted end-batch concentration after $k = 1$ is already higher than $3\sigma_{measured}$. The validation batches end with a mean relative deviation to $3\sigma_{measured}$ of -1.14% and $+2.3\%$ for $\%DE$ and IV, respectively.

As useful as it is to force the filter to provide a desired end estimate uncertainty, the resulting initialization and tuning as such does not imply a fast convergence of the end-time (t_f) prediction with the actual t_f measurement. The uncertainty of the estimate at t_f could be indeed within the defined $\pm 3\sigma_{measured}$, however, when the mis-modelling is accentuated the P matrix can be erroneous and of little use. Due to the lack of correction in the dynamics of the filter model to match the actual measurements leads to under-predictions in both state variables. Indeed, this is what happens in this system, as it can be seen for example in extraction 7 as showed in Figure 7.11. For both variables, the uncertainty is as desired at t_f, but the filter fails to converge at fast enough rate to the last target measurement.

With this initialization and tuning, we obtain the results in Table 7.2. In this table, the end-time relative prediction errors:

$$\text{relative prediction errors}^{t_f}(\%) = \left(\frac{y_{end} - \hat{x}_{k|k}^{propagation,t_f}}{y_{end}} \cdot 100\right) \tag{7.32}$$

are calculated for each k time update in the 12 batches. The behaviour of the IV and $\%DE$ filters for all extractions is identical to the one seen in Figure 7.11, with a slowly improving prediction as more data is collected. The predictions for IV are better than the ones for $\%DE$. Despite the poorer initial conditions guess $\hat{IV}_0$ provided (and P_0^{IV} is not sufficiently large to cover the measured data), the IV filter is able to recover faster than the $\%DE$ filter which has a very good initial condition estimate. This is due to a closer model dynamic to the real behaviour.

The concentration filter has a different performance in the sense that its initial prediction is much closer to the end measurements, due to a good guess for $\hat{x}_0$, and its first time-update incurs a deviation from this prediction. This is representative of an un-modelled fast diffusion stage. The $k = 1$ data are significantly under-predicted by the model, and when the time update occurs the model perceives this correction to a higher value as a difference in pectin content load rather than its dynamic. This can be seen for extraction 7 in Figure 7.12. The filter is able to recover to acceptable distances to measurement after $k = 2$ for most extraction cases, the worst case being extraction 8 where a bigger deviation for $k = 0$ is registered.

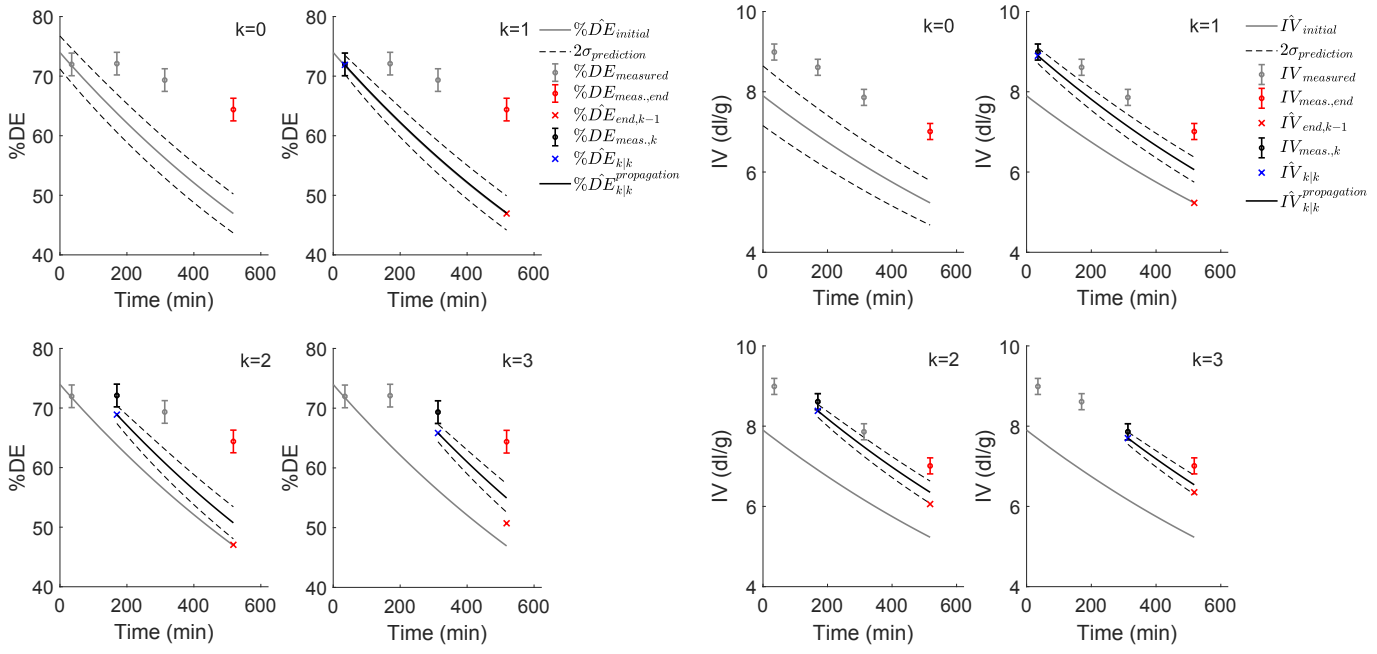

Figure 7.11: Predictor performances for $\%DE$ and IV throughout extraction 7. The figures follow the sequence of events which are triggered by acquiring data at each time k. The effect that the CD-EKF algorithm has on the new prediction until the end batch time ($\hat{x}_{k|k-1}$) is illustrated by plotting the initial simulation ($k = 0$) $\hat{x}_{initial}$ and the previous ($k - 1$) predicted endpoint estimate $\hat{x}_{end,k-1}$.

Table 7.2: Relative errors of prediction at end of extraction time ($\hat{x}_{k|k}^{propagation,t_f}$) relative to the measured output at t_f. The errors are calculated for each prediction step after the time-update (k). The initial model errors are also indicated ($k = 0$).

	Extractions											
% Relative error	1	2	3	4	5	6	7	8	9	10	11	12
$C_{pectin.bulk}$												
model (k=0)	5.65	-4.42	7.36	3.12†	-0.31†	-1.52†	2.20†	18.99	-0.15†	-1.71†	1.89†	-1.93†
k=1	24.81	6.09†	42.63	34.47	26.66	17.05	45.21	33.47	45.42	37.96	53.68	52.66
k=2	14.91	13.06	13.50	8.68	-6.86†	16.01	16.45	27.33	5.03†	-8.12†	-5.39†	-0.60†
k=3	-0.77†	5.48†	0.56†	-17.21	0.02†	-2.15†	-8.44†	10.00	-0.15†	-5.82†	-8.50†	-10.79
k=4	-8.37†		1.43†	-4.62†	-5.13†	-2.85†		3.02†	-4.45†		-7.93†	3.71†
$\%DE$												
model (k=0)	-26.72	-27.88	-28.07	-25.33	-27.70	-28.79	-27.10	-26.33	-29.58	-22.60	-29.58	-28.46
k=1	-27.58	-22.92	-28.70	-25.79	-27.70	-29.67	-26.94	-24.47	-28.59	-24.13	-28.24	-27.91
k=2	-22.94	-16.35	-23.90	-21.82	-22.66	-25.02	-21.21	-21.88	-24.08	-17.72	-28.17	-23.86
k=3	-16.20	-11.86	-17.50	-15.65	-18.75	-18.96	-14.63	-16.99	-19.22	-11.80	-21.76	-18.97
k=4	-9.63		-11.05	-9.76	-11.71	-12.73		-12.20	-11.95		-15.18	-14.22
IV												
model (k=0)	-25.80	-23.66	-23.10	-23.62	-23.45	-26.18	-25.37	-21.86	-25.52	-21.83	-24.30	-22.47
k=1	-15.01	-7.38	-13.97	-10.69	-13.34	-9.63	-13.57	-3.75	-15.85	-8.67	-15.74	-8.70
k=2	-11.34	-3.82	-10.21	-6.62	-8.78	-8.23	-9.39	-4.58	-11.77	-5.88	-9.84	-5.37
k=3	-7.59	-3.19	-6.82	-5.19	-6.08	-6.95	-6.72	-3.06	-7.82	-2.79	-7.13	-4.67
k=4	-4.23		-4.78	-3.30	-5.14	-4.81		-2.79†	-4.74		-5.22	-3.53

† Prediction falls within the last measurement $\pm 2\sigma_{measured}$ range.

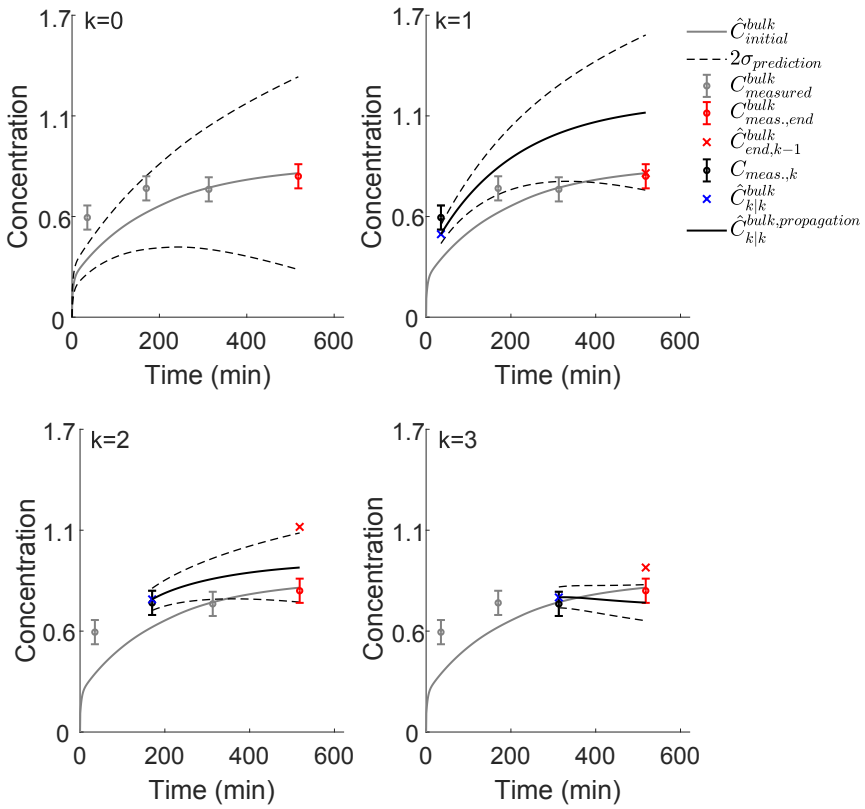

Figure 7.12: Predictor performance for Concentration throughout extraction 7. The figures follow the sequence of events which are triggered by acquiring data at each time k. The effect that the CD-EKF algorithm has on the new prediction until the end batch time ($\hat{C}_{k|k-1}$) is illustrated by plotting the initial simulation ($k = 0$) $\hat{C}_{initial}$ and the previous ($k - 1$) predicted endpoint estimate $\hat{C}_{end,k-1}$.

This section showcased the process of implementing systematically a working filter, taking into consideration the *apriori* knowledge of parametric and raw material variables uncertainty. It also highlighted how a predictors performance can be tuned, through the sum of squares minimization of tuning parameter k_Q, to render a desired prediction performance in terms of uncertainty. However, there are evident complications to circumvent if there is a goal to use the CD-EKF for improved prediction capability whilst the process is on-going. These are tackled in the following subsection.

7.3.2 Implementation corrections

In this subsection, the shortcomings of the filters are addressed in particular for the IV and $\%DE$ cases. The concentration model is however adequate to a certain degree, this can be seen in Table 7.2, where the initial model predictions of the concentrations at t_f are far superior to the other models. However, there is an issue of a poorly characterized rapid-diffusion which corrupts the filter performance and, given the model's poor observability, it would be more adequately tackled in an off-line parameter estimation to the production scale data and then posthumous filter implementation. The initial state and parametric deviation of $\%DE$ and IV are deemed as problems that can be tackled with in-process data that can still be useful for on-line correction.

7.3.2.1 Recovering from poor parameters

The $\%DE$ filter convergence to the end-point data suffered from a poorly adjusted $k_{de-esterification}$ parameter estimate. The real de-esterification kinetics were in reality much slower than what the parameter, estimated in a different context, anticipated. The correction in this dynamics can be made by augmenting the state vector with the parameters desired to be corrected. For the $\%DE$ system case (same for IV), the only system structures where observability is guaranteed (non-zero determinant of the observability matrix) is when only one parameter is being corrected.

If we take into consideration the most significant parameter to estimate, which we can take from the sensitivity results in chapter 5 (see Figure 5.11), the extended

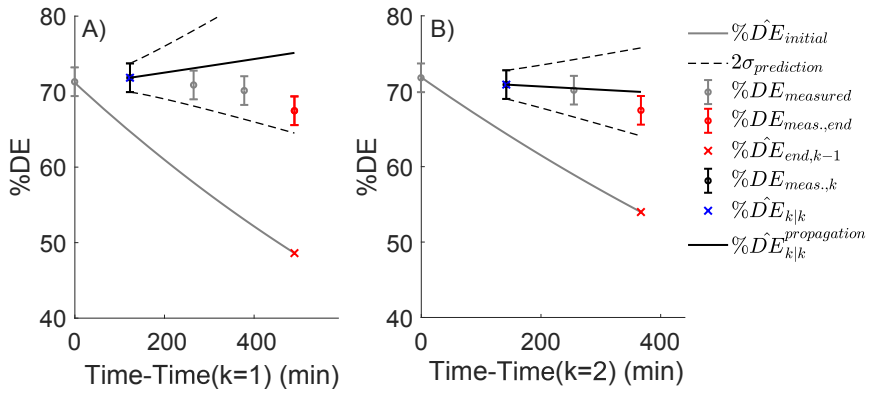

Figure 7.13: Joint $\%\hat{DE}$ state and $\hat{\alpha}_{DE}$ estimation performance in extraction 1 for A) inclusion of all data-points and B) Exclusion of data-point k=1. The filter algorithm is initialized from k=1 for A) and from k=2 for B). The initial prediction is plotted to showcase the difference between pure model simulation and the CD-EKF implementation.

state variable becomes:

$$\frac{dx_{\%DE}}{dt} = \begin{bmatrix} \%\dot{D}E \\ \dot{\alpha}_{DE} = 0 \end{bmatrix} \tag{7.33}$$

Here the $\hat{x}_0$ for the augmented parameter can be defined as the estimate from the parameter estimation in Chapter 5, $\hat{\alpha}_{DE,0} = 8.46$ (1/(mol s)). The measurement noise covariance R matrix does not change and the process noise covariance can be defined as such (Grewal and Andrews 2008):

$$Q(t)_{augmented} = \begin{bmatrix} Q(t)_{states} & 0 \\ 0 & 0 \end{bmatrix} \tag{7.34}$$

where $Q(t)_{states}$ corresponds to the process noise covariance matrix developed for the previous cases of sole state estimation. Note that the parameter values will be changed throughout the course of time and these updated values will be inputs in $Q(t)_{states}$.

For the first extraction, if the filter begins at $k = 1$ rather than $k = 0$, the $\hat{x}_0$ estimate for the state variable can be replaced by $x_{meas.,k=1}$ and its initial estimate error variance can be replaced by the known measurements variance ($\sigma^2_{prediction}$). This avoids an erroneous initial state estimate and helps ensure we identify the desired kinetic parameter in good time to provide information for the following batches. Regarding the initial estimate error covariance matrix (P_0), it will affect only the transient phase. By defining a large P_0 for the parameter state, we are forcing a fast recovery of this parameter. For this case, $P_{0,\hat{\alpha}_{DE}} = 10^5$ is chosen.

It is important in this case to have a data rejection safeguard to ensure that the parameter estimation is achieved correctly. In Figure 7.13 it can be seen the effect that a wrong sequence of data-points has on the posterior estimate of the system dynamic. In Figure 7.13 A) all data-points of extraction 1 and considered and it can be seen that $\%DE(k = 1) < \%DE(k = 2)$. The filter assumes this slope signal change as the real behaviour which leads to a wrong estimate and a poorer prediction. For this case, the filter never recovered to an acceptable performance and parameter estimate. This is also what would happen if we started the algorithm at $k = 0$ and had a $\%\hat{D}E_0$ below $k = 1$. If the data rejection condition $\%DE(k = 1) > \%DE(k = 2)$ is ensured, then the k=1 data-point would be discarded and the k=2 would become the starting point for the algorithm. This is showcased in Figure 7.13 B), where the prediction to t_f can be seen to improve significantly from the initial prediction.

7.3.2.2 Recovering from poor initial conditions guess

The previous IV filter benefited from having better parameter estimates, even though the kinetics could still be improved. However, the initial state estimate (x_0) was underestimated. In the context of the sequential extraction operation, it would benefit the production to provide an updated initial state estimate ($x_0^{new,i}$ and $P_0^{new,i}$) for the

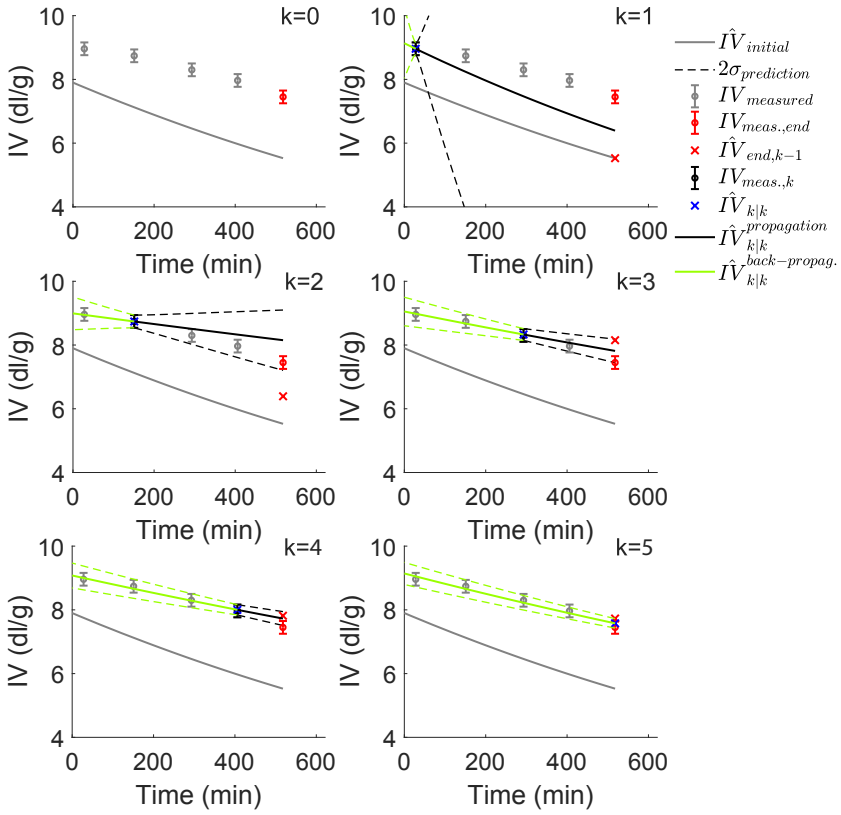

Figure 7.14: Joint $\hat{IV}$ state and $\hat{\beta}_{IV}$ estimation with the forward and backward prediction propagations at each data-point k for extraction 1. The effect that the CD-EKF algorithm has on the new prediction until the end batch time ($\hat{IV}_{k|k-1}$) is illustrated by plotting the initial simulation ($k = 0$) $\hat{IV}_{initial}$ and the previous ($k-1$) predicted endpoint estimate $\hat{IV}_{end,k-1}$.

initialization of the forthcoming batches. Since there are no actual $k = 0$ data, there is a need for extrapolation from the current state estimates and error covariances. Considering the sensitive parameter augmented IV filter, as was done for $\%DE$, the state space vector is the following

$$\frac{dx_{IV}}{dt} = \begin{bmatrix} \dot{IV} \\ \dot{\beta}_{IV} = 0 \end{bmatrix} \tag{7.35}$$

Here a large P_0 for the parameter state is again defined and, similarly to the previous section, $P_{0,\hat{\beta}_{IV}} = 10^5$ is chosen. The same data rejection is employed to

avoid extrapolating an erroneous initial estimate. The extrapolation of $k = 0$ can be made from the current k through backwards propagation of both the state and error covariance. For simulations where time is evolving reversely, the sign on the dynamic coefficients in the systems (7.34) and (7.35) changes. This also changes the Riccati equation for the following expression:

$$\frac{dP_{back}}{dt} = -A(t)P - PA(t)^T + Q(t) \tag{7.36}$$

In Figure 7.14, it is possible to assess the different back-propagation at the different k measurements. It is easy to assess that from $k = 2$ on, the newer $\hat{\beta}_{IV}$ estimates provide us a different dynamics, which is then reflected on the new predictions for IV_0 derived from the backwards propagation.

This is repeated for all the extractions and the resulting IV_0 estimates from back-propagation at different measurements k can be seen in Figure 7.15. In all cases, it is observed that there is a good agreement for the different estimates at an extraction level. Considering all the extractions, it is also visible that the IV_0 follows a random distribution around the overall mean. For a given extraction, all the estimates are on the same side of the IV_0 mean distribution. The variation between extractions represents the expected natural variability within one lot of the same raw material.

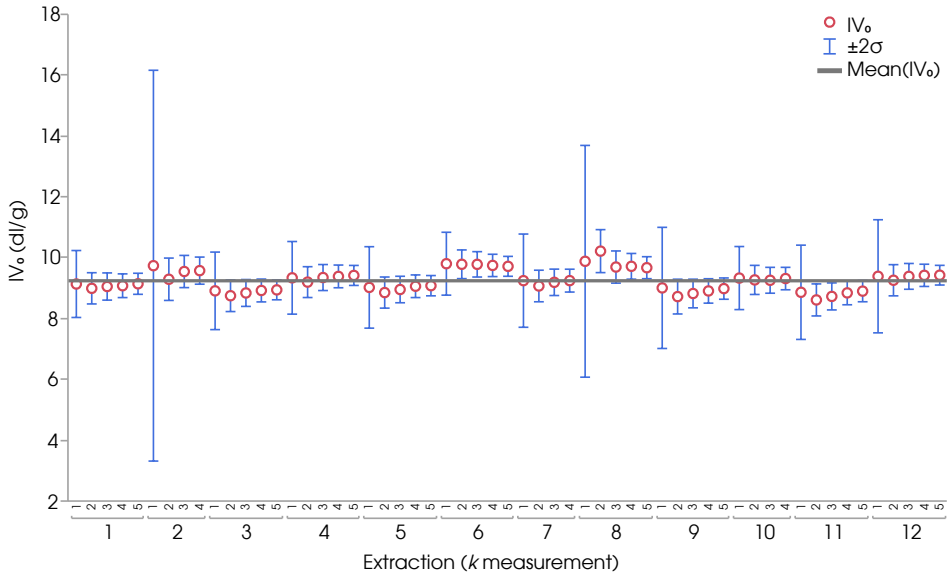

Figure 7.15: Predicted $\hat{IV}_0$ from backwards propagation at the different measurements k for each extraction. The 2σ confidence interval derived from the propagated $P_{back,t=0}$ is shown for each backward prediction and the mean line of all $\hat{IV}_0$ estimates is plotted.

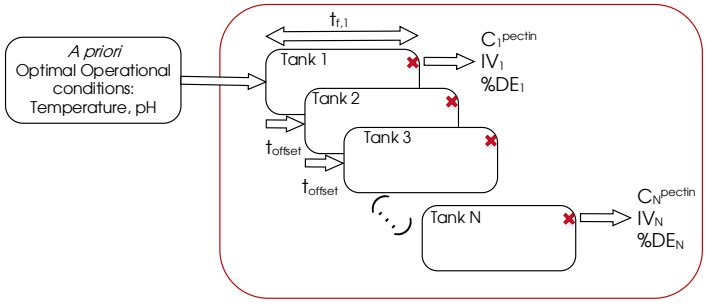

Figure 7.16: Representation of the successive extraction series in this case study

The estimates for $k = 1$ reveal a higher uncertainty, given that we initialize the filter with a large estimate error covariance matrix and at $k = 1$ the filter has been through less measurement update iterations, which lower the P matrix values. For extractions 2 and 8 this uncertainty is relatively higher compared to the rest, due to the fact their $k = 1$ measurements being later in time. This implies a longer back-propagation in time, which leads to higher uncertainty.

7.3.2.3 Combination in the context of an extraction series

The production runs the extractions series in a successive fashion (i.e., extraction 2 starts after extraction 1 and so forth). This allows for an operation which learns throughout the extent of the past extractions to enable better predictions in the following batches. A pictorial representation of the nature of the operation is shown in Figure 7.16.

The scenario where the first 6 batches (calibration) are used to provide better initialization to both state ($\hat{x}_0$) and sensitive parameter ($\hat{p}$) to the following 6 simulations (validation) is devised. The approach is performed as an on-line algorithm within the batches rather than an off-line estimation exercise to avoid waiting for a full extraction to be completed to have a correct trajectory of our states. This could imply bad start-ups (erroneous "optimal" conditions) in several other extractions. It is more advantageous to improve the predictions on an ongoing basis. However, the first extraction has to start-up based on the series *apriori* information.

In Table 7.3 the validation batches are assessed for their prediction capability, given the information provided from the previous calibrations batches. This is assessed via the relative error of the end batch prediction to the respective batch t_f measurement. This is information that transits from the six calibration batches can be passed in different forms. Four different scenarios are compared in Table 7.3. This scenarios are permutations of using a mean of means for all the estimates in the six calibration extractions, or simply using a mean for latest $\hat{x}_0$ estimates across the different calibration extractions. The outputs show that scenario B, which makes the

mean of $\hat{x}_0$ and $\hat{p}$ using all k_{all} measurement estimates, provides smallest mean of relative errors for both variables: $\%DE$ - 1.43% and IV - 1.47%.

In an attempt to assess how the number of validation batches affects the performance of the improved model from the start of the process, the analysis of varying the number of calibrations extractions from 1-6 is done and summarized in Table 7.4. It is possible to assess that the performance, even with just one calibration batch, is far superior to the observed in the previous subsection (see Table 7.2). This becomes apparent by the large number of predictions which fall within the $\pm 2\sigma_{measured}$ range of the measured output at t_f. The $\%DE$ filter appears to necessitate lesser batch to achieve enhanced performance, whilst IV shows a monotonically improving behaviour with the increasing amount of calibration extractions. The same can not be said for $\%DE$. This analysis is also very dependent on the similarities among batches, and it weighs heavily on extraction 1, which could be non-representative. A better approach would be to perform this in a combinatorial fashion for the six calibration extractions.

Table 7.3: Relative errors of prediction at end of extraction time ($\hat{x}_{k|k}^{propagation,t_f}$) relative to the measured output at t_f. The errors are calculated for the initial model simulations ($k = 0$) with the updated initializations derived from the six calibration batches. Four different scenarios (A-D) for the calculation of the mean updated initial state estimates ($\hat{x}_0$) and the respective augmented parameters ($\hat{p}_0$) from the estimates of the calibration extractions 1-6 are shown for each model.

% Relative error		Extractions						Initialization	
for model (k=0)		7	8	9	10	11	12	$\hat{x}_0$	$\hat{p}$
$\%DE$								%	l/(mol s)
	A	3.32	-0.13	0.61	1.53	0.93	4.48	73.6	1.14
	B	-0.08	-2.96	-2.78	-1.00	-2.50	0.76	73.9	1.91
	C	3.76	0.29	1.04	1.96	1.36	4.93	73.9	1.14
	D	-0.50	-3.37	-3.19	-1.41	-2.92	0.33	73.6	1.91
IV								dl/g	s^{-1}
	A	-0.45	4.34	-0.59	1.44	0.98	5.57	9.31	24.9
	B	-0.85	3.92	-0.99	1.00	0.57	5.16	9.26	24.5
	C	-0.96	3.81	-1.10	0.92	0.47	5.04	9.26	24.9
	D	-0.35	4.45	-0.49	1.52	1.09	5.70	9.31	24.5

[A] Mean of $\hat{x}_0$ and $\hat{p}$ using only the last k_{end} measurement estimates.
[B] Mean of $\hat{x}_0$ and $\hat{p}$ using all k_{all} measurement estimates.
[C] Mean of $\hat{x}_0$ using all k_{all} measurement estimates and mean of $\hat{p}$ using only the last k_{end} measurement estimates.
[D] Mean of $\hat{x}_0$ using only the last k_{end} measurement estimates and mean of $\hat{p}$ using all k_{all} measurement estimates.

Table 7.4: Relative errors of prediction at end of extraction time ($\hat{x}_{k|k}^{propagation,t_f}$) relative to the measured output at t_f. The errors are calculated for the initial model simulations ($k = 0$) with the updated initializations derived from the previous number of calibration batches indicated per row.

% Relative error for model (k=0)		Extraction 2	3	4	5	6	7	8	9	10	11	12	Absolute mean error
$\%DE$	N cal												
	6						-0.08†	-2.96	-2.78†	-1.00†	-2.50†	0.76†	1.68
	5					-4.62	-0.39†	-3.25	-3.09	-1.29†	-2.81†	0.43†	2.27
	4				0.59†	-3.59	0.74†	-2.25	-1.97†	-0.34†	-1.69†	1.63†	1.60
	3			-1.28†	0.65†	-3.48	0.81†	-2.09†	-1.92†	-0.12†	-1.64†	1.65†	1.51
	2		-0.37†	-0.71†	1.41†	-2.83	1.55†	-1.51†	-1.17†	0.38†	-0.88†	2.48†	1.33
	1	-0.56†	0.49†	0.02†	2.33†	-2.04†	2.45†	-0.77†	-0.28†	1.04†	0.02†	3.47	1.22
IV													
	6						-0.85†	3.92	-0.99†	1.00†	0.57†	5.16	2.08
	5					-1.50†	-0.51†	4.28	-0.65†	1.02†	0.92†	5.74	2.09
	4				2.50†	-1.27†	-0.27†	4.54	-0.41†	1.32†	1.16†	5.97	2.18
	3			2.45†	2.56†	-1.22†	-0.22†	4.59	-0.36†	1.25†	1.21†	6.10	2.22
	2		3.22	3.16	3.28	-0.51†	0.49†	5.34	0.35†	2.17†	1.94†	6.73	2.72
	1	6.33	6.07	6.27	6.35	2.40†	3.40	8.43	3.28	3.88	4.88	10.68	5.63

† Prediction falls within the last measurement $\pm 2\sigma_{measured}$ range.

7.4 Concluding remarks

In this chapter, it was shown how through the use of state estimators it is possible to incorporate the model predictions with in-process measurements to obtain an improved predictor. A systematic approach to constructing this predictor with a desired performance has been proposed. However, this would require a model which would be adequate to the specific case and in reality, the offset between the measured outputs and the filter model is unavoidable. These challenges in applying these algorithms were highlighted, namely the recovery from poor initial state guess and poor model parameter assumptions. Both situations are common in the production environment and a solution to tackle both have been presented. State space augmentation with a sensitive parameter combined with backwards propagation showed a substantial improvement to the standalone CD-EKF. This provides the operational strategy the systematic basis for reconciling the filter model with in-process data, leading to a forecast correction.

Part III
Implementation and Practical Considerations

CHAPTER 8

Integrated Operation

This chapter comprises the combination of the different tools developed in Part II and it puts them into the context of the operational strategy proposed in Chapter 3, with an illustration for the pectin extraction case. The critical material attributes of a lime peel are predicted with the FT-NIR. After defining a product quality profile, model-based optimization is performed. The model is then used as a predictor, and it is corrected with in-process samples coming from the production. Throughout this chapter, the concentration results are scaled to the +95% confidence interval at t_f resulting from the Monte Carlo simulation of the normal operating conditions (NOC) range (see Figure 7.4 B).

8.1 Series definition

The operational strategy begins prior to the start-up of the extraction series, with the definition of the process conditions at which the series should operate given a desired quality profile of the product and the raw material used. Figure 8.1 zooms in this step. Two of the developed tools will be needed in this step. Firstly the raw material assessment through FT-NIRS has been made. This peel was added in the external validation set of the models developed in Chapter 4. The prediction for $\%DE_0$ compared to its laboratory measured value can be seen in Figure 8.2. The rest of the predicted values and the uncertainty associated with the PLS models can be seen in Table 8.1. It is seen that the predicted values are in good agreement with the measured values in the lab.

Some calculations are needed since the inputs to the model in terms of concentration of readily available pectin and protopectin in the peel and the predicted values have their reference with the extracted pectin in solution. Therefore with a simple transformation:

$$C^0_{pectin,peel} = \frac{C^0_{pectin} \cdot 0.15\text{l}}{3\text{g} \cdot \frac{1}{\rho_{peel}}\text{l/g}} = 123\text{kg/m}^3 \tag{8.1}$$

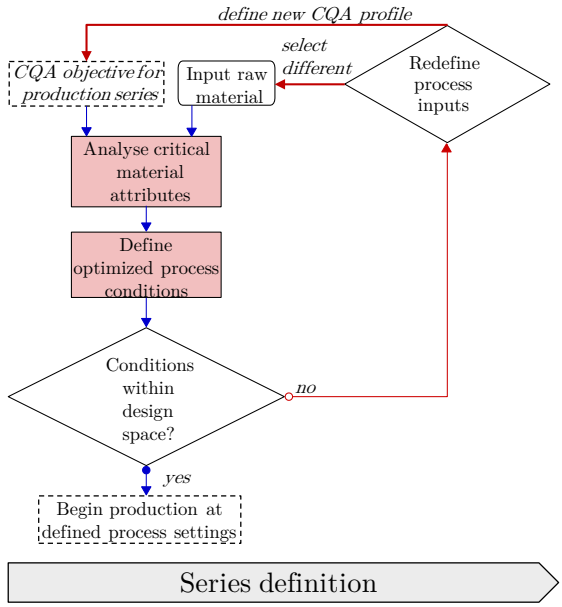

Figure 8.1: Series definition step of the operational strategy. A requested quality objective and an assessment of the raw material are necessary to define the conditions at which the process should run.

$$C^{total} = \frac{Yield \cdot \frac{1}{100} \cdot 3\text{g}}{3\text{g} \cdot \frac{1}{\rho_{peel}}\text{l/g}} = 273\text{kg/m}^3 \tag{8.2}$$

$$C^0_{protopectin,peel} = C^{total} - C^0_{pectin,peel} = 150\text{kg/m}^3 \tag{8.3}$$

here the 3 g and 0.15 l correspond to the mass of peel and solvent used in the laboratory reference, respectively. To convert mass to volume, the ρ_{peel}=1030 l/g

Table 8.1: Critical material attributes estimates from PLS-R models and their respective uncertainty. The estimates are the mean value of the triplicate spectra predictions. In the parentheses are the values obtained in the lab analysis

	$\hat{u}_{peel}$	σ_{PLS-R}
$\%DE_0$	74.0 (74.7)	1.05
IV_0 (dl/g)	8.46 (8.00)	0.58
C^0_{pectin} (g/l)	2.39 (2.55)	0.20
%Yield	26.5 (28.8)	1.57

is considered. This arithmetic operations require the error propagation from the corresponding σ_{PLS-R}. At the end the uncertainties of the concentration model inputs are $\sigma_{C^0_{pectin,peel}} = 10.3$ kg/m^3 and $\sigma_{C^0_{protopectin,peel}} = 19.2$ kg/m^3.

As for the CQA objective, for the demonstration purposes, we can select the end-quality $\%DE(t_f)$ and $IV(t_f)$ at the end of the first extraction as the desired quality. Subsequently, the results of the model-based optimization for pH and temperature with the conditions at which the process was actually processed can be compared. The last time point of the extraction was 9h37min ($t_f = 577$ min) and registered $\%DE(t_f) = 63.4$ and $IV(t_f) = 5.75$ dl/g. The time is an optimized variable in the routine, but here is kept in a tight bound of 15 min around t_f. Temperature and pH and bounded by the NOC values. The constraints around the final $\%DE$ and IV are set at $\pm 3\sigma_{measured}$. Given the outcome of the model simulations in Chapter 7 and this peel being also a lime, the estimated parameters presented in Table 7.3 for scenario B are used. The worst case optimization scenarios, based on the raw material uncertainty yields the conditions shown in Table 8.2. In reality the extraction was run at milder conditions T= 65.5℃ and average pH=1.43. These values are not far off from the optimization results and indicate that our model has an expected good agreement with the plant data and it would thus, in combination with the NIRS, provide a very acceptable initialization for the extraction series.

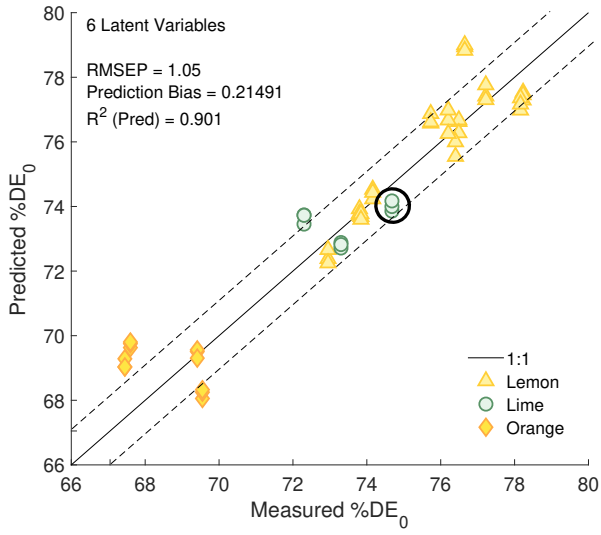

Figure 8.2: The selected peel highlighted (black circle) in the predicted vs measured $\%DE_0$ plot presented in the previous Chapter 4.

Table 8.2: Result of robust optimization of pectin extraction from the given lime peel and the desired for the desired target $\%DE(t_f) = 63.4 \pm 0.95$ and $IV(t_f) = 5.75 \pm 0.1$.

T (℃)	pH	Batch time (min)	Perturbation	$C_{pectin}(t_f)$	$IV(t_f)$ (dl/g)	$\%DE(t_f)$
70	1.38	577	$\mu + \sigma$	0.98	6.37	64.17
			$\mu - \sigma$	0.79	5.56	62.4

8.2 Forecast correction

As the first extraction starts the filter can run solely as state estimator with the tuning parameters k_Q defined such that the uncertainty propagates as desired, as shown in section 7.3.1. However, the assessment of the deviation to the measurements and the dynamics correction to update the ongoing extraction, and more importantly provide better initialization to the next, can be performed in parallel and this is pictured in Figure 8.3.

As initial state estimates ($\hat{x}_0$) the u_{peel} variables obtained through the PLS models can be considered. The associated σ^2_{PLS-R} is then the initial estimate error P_0 diagonal values. If we consider the previously defined measurement noise covariance as before (see Table 7.1) and the same $k_{Q,C} = 1$, $k_{Q,\%DE} = 5.12 \cdot 10^5$ and $k_{Q,IV} = 5.67 \cdot 10^4$ the filter performance in Figures 8.4-8.5 is obtained for the first extraction.

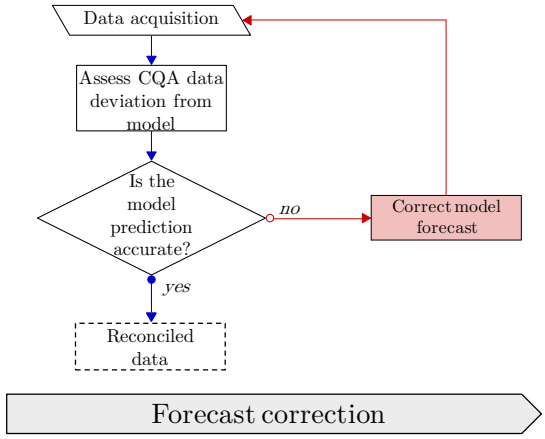

Figure 8.3: Forecast correction step of the operational strategy. As data is acquired, the model is compared and corrected if necessary.

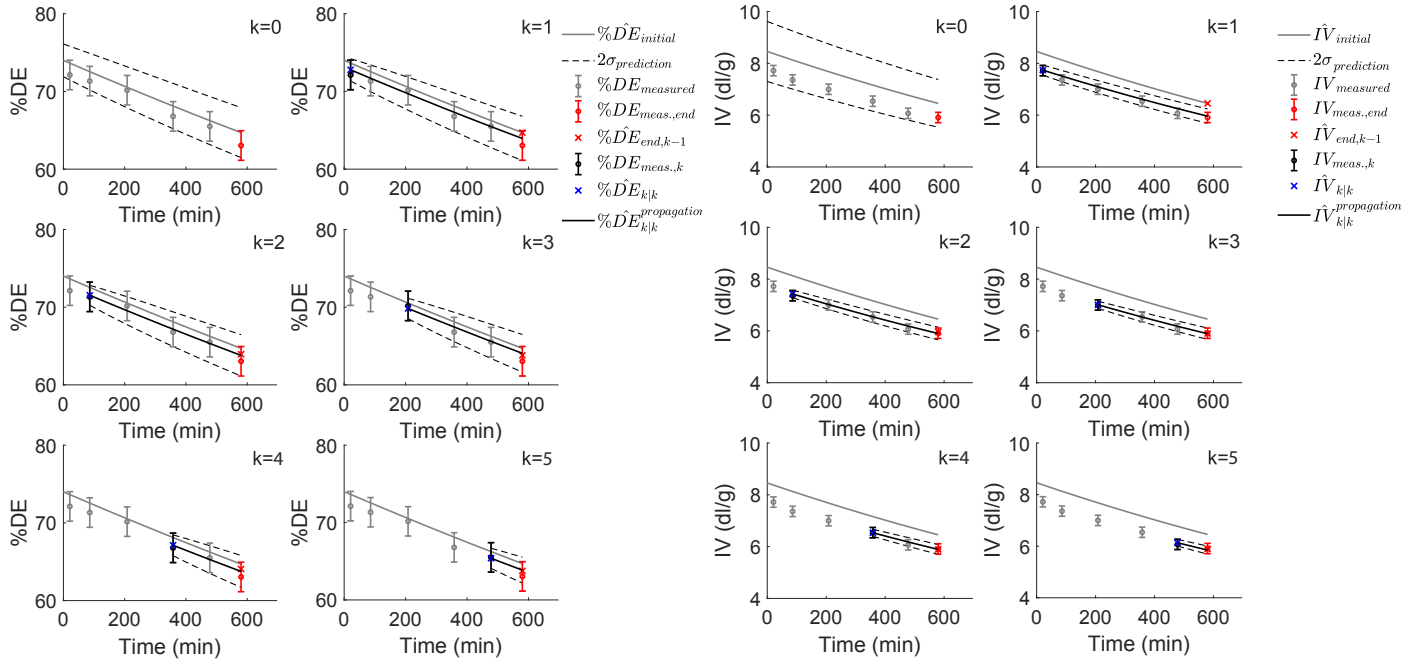

Figure 8.4: Predictor performances for $\%DE$ and IV. The figures follow the sequence of events which are triggered by acquiring data at each time k. The effect that the CD-EKF algorithm has on the new prediction until the end batch time ($\hat{x}_{k|k-1}$) is illustrated by plotting the initial simulation ($k = 0$) $\hat{x}_{initial}$ and the previous ($k - 1$) predicted endpoint estimate $\hat{x}_{end,k-1}$.

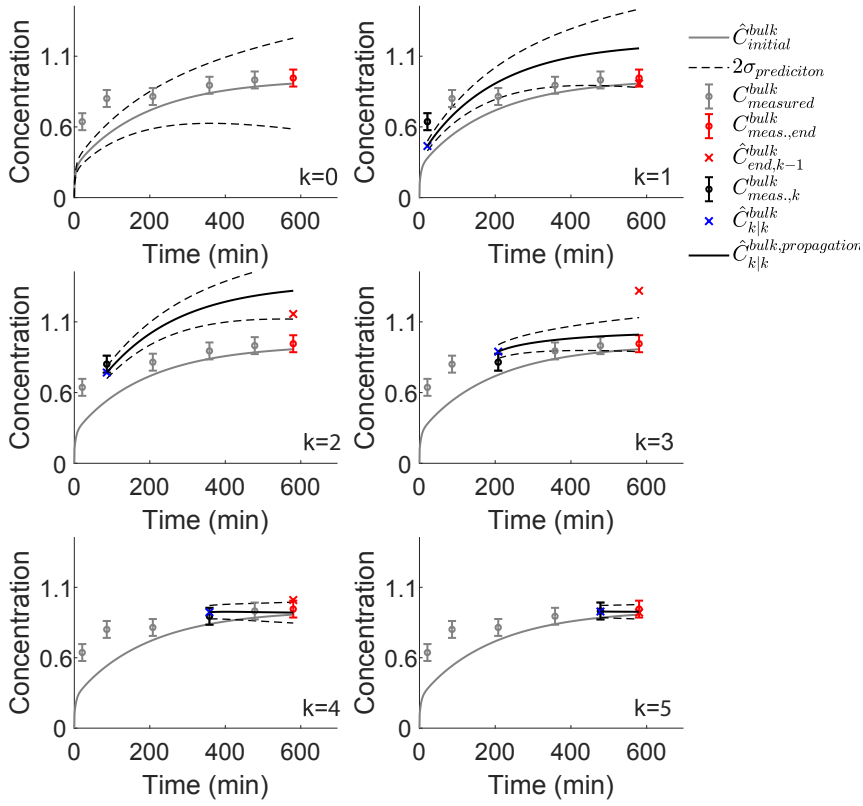

Figure 8.5: Predictor performance for Concentration. The figures follow the sequence of events which are triggered by acquiring data at each time k. The effect that the CD-EKF algorithm has on the new prediction until the end batch time ($\hat{C}_{k|k-1}$) is illustrated by plotting the initial simulation ($k = 0$) $\hat{C}_{initial}$ and the previous ($k - 1$) predicted endpoint estimate $\hat{C}_{end,k-1}$.

The filter has a good performance for both $\%DE$ and IV due to the correct initial state estimates (and associated uncertainty) from the PAT application and the correct sensitive parameters coming from the augmented state estimation in Chapter 7 with a similar fruit at the same scale. The concentration suffers from the same problem as seen in Chapter 7, with a poor portrayal of the initial washing phase. However, given the correct $C^0_{pectin,peel}$ and $C^0_{protopectin,peel}$ initial states were provided through the PLS-R model, the model can predict the concentration at t_f at the beginning of the extraction with accuracy. The model pitfalls are well known; thus a high degree of trust could be given to the initial simulation prediction.

Table 8.3: Relative errors of prediction at end of extraction time ($\hat{x}_{k|k}^{propagation,t_f}$) relative to the measured output at t_f. The errors are calculated for each prediction step after the time-update (k). The initial model errors are also indicated ($k = 0$).

% Relative error	Extractions		
	1	2	2 updated
$C_{pectin.bulk}$			
model (k=0)	3.98	-4.62	
k=1	42.51	24.87	
k=2	49.34	44.29	
k=3	34.05	7.62	
k=4	8.64	-2.93	
k=5	-4.75	-2.27	
k=6	-0.67		
$\%DE$			
model (k=0)	2.71	2.63	-0.53
k=1	2.14	1.44	-1.18
k=2	1.91	1.20	-1.11
k=3	1.65	1.63	-0.3
k=4	1.21	1.13	-0.27
k=5	0.47	1.29	0.35
k=6	-0.12		
IV			
model (k=0)	13.94	9.22	-2.03
k=1	-1.48	0.90	4.71
k=2	-1.75	-0.18	3.57
k=3	-1.34	-0.31	2.99
k=4	-0.18	-0.29	2.26
k=5	0.30	-1.04	0.75
k=6	1.42		

By employing the joint state and parameter estimation together with the backward propagation, as developed in Chapter 7, an improved guess for the initial states, $\%DE_0 = 73.4\%$ and $IV_0 = 7.27$dl/g, and parameters, $\alpha_{DE} = 2.43$ l/(mol s) and $\beta_{IV} = 6.94$ s^{-1}, is provided to the second extraction. This results in a fairly improved initial prediction as can be seen in Table 8.3. However, for IV the following predictions (from k=1 on) are seemingly worse than the original filter. The updated filter for IV can be seen in Figure 8.6. Even though the errors are worse the performance is still very positive. The majority of the end estimates fall in the end-point $\pm\sigma_{measured}$ range. Thus a slightly less precise but still accurate filter is obtained in exchange for a far more accurate ($\times 4.5$) initial prediction.

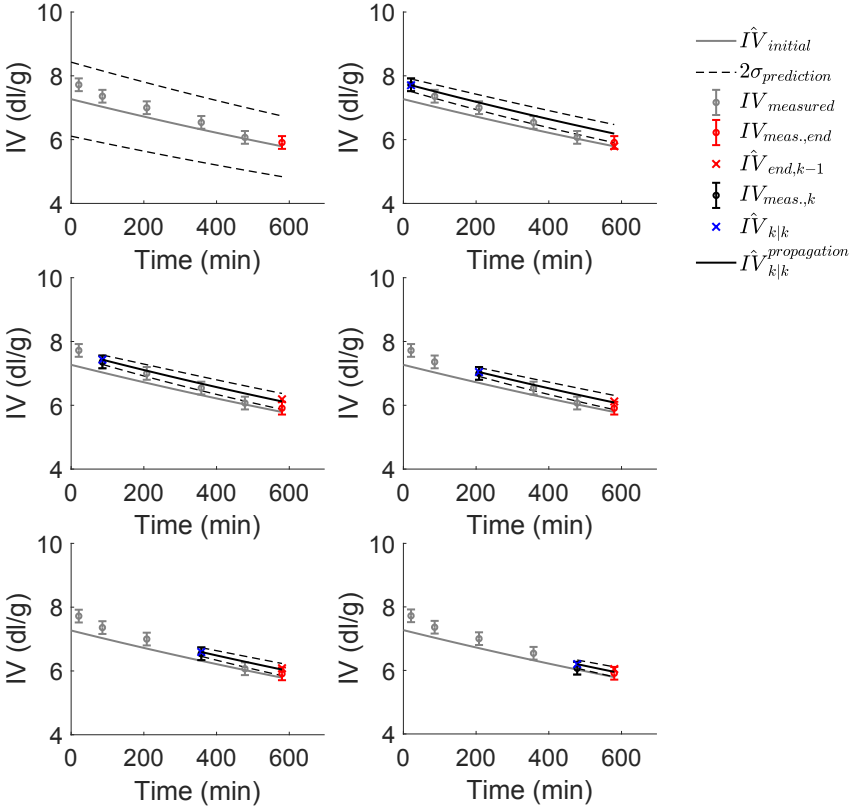

Figure 8.6: Predictor performance for IV. The figures follow the sequence of events which are triggered by acquiring data at each time k. The effect that the CD-EKF algorithm has on the new prediction until the end batch time ($\hat{IV}_{k|k-1}$) is illustrated by plotting the initial simulation ($k = 0$) $\hat{IV}_{initial}$ and the previous ($k-1$) predicted endpoint estimate $\hat{IV}_{end,k-1}$.

8.3 Process conditions correction

As the deviation from the target quality profile is assessed with re-assured confidence in our model, a decision of maintaining the current process conditions for the posterior iterations or re-defining the optimal conditions with the support of the more confident model is made. With the filter model from the previous subsection, the confidence established is quite high. We could confidently say that the process will finish at the simulated t_f given the necessary in-process measurements were provided.

At this point, an assessment of the deviation to the initial desired quality target needs to be made. If indeed the extraction conditions are leading to outputs which

unacceptably are out-of-specification, then through a new iteration of the model-based optimization, this time around with improved model dynamics and u_{peel} parameters from the forecast prediction correction, the new optimal conditions can be re-defined. This is pictured in Figure 8.7.

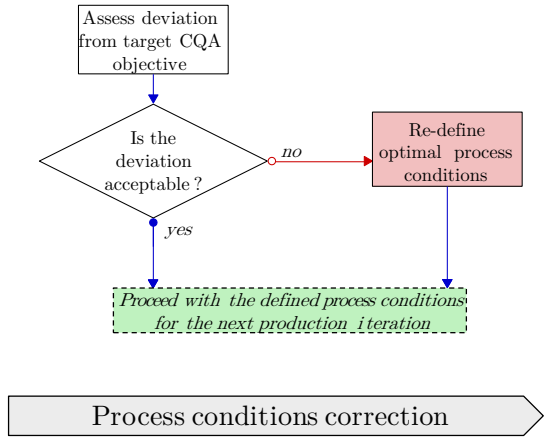

Figure 8.7: Process conditions correction step of the operational strategy.

CHAPTER 9

Conclusion

In this thesis, an operational strategy for bio-based operations was proposed, and the development of the necessary tools was showcased for an industrial case-study focusing on the performance monitoring of a pectin batch extraction process. This was made possible through the collaboration with CP Kelco.

A brief overview of bio-based production and the concepts of measuring, modelling, monitoring, and control set the tone for the goal definition. With these concepts defined, a clear motivation and objectives on how to apply these methods to tackle the biological raw material induced disturbances was stated. A more detailed review of the developments in solid-liquid extraction is given, highlighting the current practices. Given the perspective of the review, an operational strategy is proposed, inspired by Quality-by-Design foundations, with the intention of enhancing process performance and reduce batch-to-batch discrepancies induced by the processing of biological feedstock. This roadmap comprises the combination of monitoring and optimization components. The tools developed for the case study were a combination of both data and model-driven soft-sensors.

Raw material characterization approaches reliant on near-infrared spectroscopy were assessed. The coupling of a spectroscopic tool with chemometric modelling revealed that this application is a valid alternative to the current reference methods. The performance of the PLS models faired quite positively, proving to be equal or less uncertain when compared with the typically used expert-knowledge discrimination (classification of peels in fruit types). This opens the possibility to transition from the inefficient production feedback as the source of raw material quality feedback, to this type of faster and less laborious technology. This application can prospectively limit the execution of reference method tests to a role of model re-calibration and maintenance. More importantly, it provided estimates of the critical material attributes (raw material specific inputs) $u_{peel} = \left[C^0_{pectin},\ C^0_{protopectin}, IV_0,\ \%DE_0\right]$, which were targeted to a specific raw material and were used in further process optimization.

The relationship between the critical process conditions and the output dynamics was mapped through different types of modelling approaches. Empirical models highlighted the most influential parameters (and combinations) for the different process outputs. Kinetic modelling showed a good fit to the data and the models were flexible to variations of temperature and pH conditions. A model developed by Andersen et al. (2017), reliant of diffusion-based phenomena, was selected to be analysed for

parametric local sensitivity analysis. The analysis revealed that the most significant parameters for all variables (pectin concentration, $\%DE$ and IV) were related to the critical material attributes. It also highlighted that the parameter estimation previously performed was slightly ill-conditioned, and could benefit from parameter reduction. A new parameter estimation was performed based on this assessment and the derived model was assessed for the impact of raw material uncertainty through Monte-Carlo. The simulation performance matched well with the measured variabilities of pectin concentration, $\%DE$ and IV across the different fruits.

The model was used to provide optimal process conditions (i.e., temperature, pH and time) considering not only the variability of the raw material (i.e., three different types of fruits) but also meeting the desired quality criteria constraints for jam and jelly products. It was seen that taking the raw material into consideration in the optimization routine provided fewer quality constraints violation even if at the expense of the amount pectin extracted. Using more conservative bounds and better informed critical material attributes (i.e., lab reference method, NIRS.) would lead to fewer constraint violations.

A systematic approach to constructing a model-based predictor through the use of continuous-discrete Kalman filters was proposed and it was possible to incorporate the model predictions with in-process measurements to obtain an improved predictor. A strategy to recover from poor initial state guess and poor model parameter assumptions allowed a correction to the model initializations throughout the production series, which improved the predictor performances significantly at the different $t = 0$. The strategy included joint state space and sensitive parameter estimation combined with backwards propagation. The impact of the using a different number of batches to perform these corrections was assessed and it revealed that improvements can be made from using just one batch, but more robust approaches would require a larger calibration set.

A general limitation of the work is the lack of designed proof of the proposed methodology. The tools developed were assessed and showed applicable with satisfactory performance separately and while the combination of these tools was still tested for a specific peel in Chapter 8, the data was of happen-stance nature and it was not possible to pursue the operational guideline as if the production was being conducted *live*. However, the adequacy of the combination of the developed different tools was still attested through the incorporation of the different results which yield a sound and concordant performance.

9.1 Future perspectives

Throughout the work which has been developed for this thesis, several open questions have been raised which potential projects have not been pursued in the interest of time.

A few important application-specific questions were not addressed in the tool development in this case-study. One of them is the particularities of working with in-

frequent time-delayed samples. The samples should be as little as possible in number, with an appropriately feasible frequency, such that the model simulation is corrected as fast as possible. A possible solution for the sample positioning in the batch, time-wise, could be given by the solution of an optimization problem formulated in respect to a CD-EKF implementation (where $P(t_f)$ is the estimate error covariance matrix from the CD-EKF at the end of the extraction batch)

$$\begin{aligned} \underset{z}{\text{minimize}} \quad & J = det(P(t_f)) \\ \text{subject to} \quad & \text{constraints for the distance between samples taking} \\ & \text{into account logistics of sampling frequency, sampling} \\ & \text{time and delay time} \end{aligned} \tag{9.1}$$

where z would be in this case the vector of time-points for samples in the total time span of an extraction.

Another practical issue was that the measurements were assumed in this case-study as to have no autocorrelation, albeit in reality it some degree of correlation may exist since the analytical reference utilizes the same process equipment. This was outside of the scope of this study. However it should be assessed.

Regarding the chemometric models, it is unavoidable the fact that atypical peel samples which are outside of the calibration set will fail and thus there is a need to devise the appropriate design space for these materials and ensure that the raw materials used for production are comprised. Model update and re-calibrations are an important part of the PAT tool life-cycle. A possible extension from this work is the opportunity to recalibrate the models with the feedback information provided by the updated initial state estimate stemming from the predictor correction with in-process measurements.

A direct outcome of this thesis would be the development of an integrated dashboard which has embedded the different tools that were developed. As raw material is characterized, this information can be stored and made available for the production planning teams to assess and define which raw materials to use, depending on the product quality target at that moment. This would provide visualization ease to production planning teams and make it possible for them to define realistic product quality targets for the existent raw materials. This tool could additionally be extended to be used on the *floor* level and offer updated predictions and suggest new optimal conditions if necessary throughout the production series. In the studied case study system, the manipulated variables could only change between batches, however, for systems that allow it, an in-process control component could be added to the operational procedure.

Bibliography

Abidin, Z. Z., Awang Biak, D. R., Mohamed, H., and Harun, M. Y. (2013). "Solid–Liquid Extraction in Biorefinery". In: *Separation and Purification Technologies in Biorefineries*. Wiley-Blackwell. Chap. 13, pp. 351–374.

Andersen, N., Cognet, T., Santacoloma, P., Larsen, J., Armagan, I., Larsen, F., Gernaey, K., Abildskov, J., and Huusom, J. (2017). "Dynamic modelling of pectin extraction describing yield and functional characteristics". In: *Journal of Food Engineering* 192, pp. 61–71.

Antony, J. (2014). *Design of Experiments for Engineers and Scientists (2nd Edition)*. eng. Oxford: Elsevier.

Asl, A. H. and Khajenoori, M. (2013). "Subcritical Water Extraction". In: *Mass Transfer - Advances in Sustainable Energy and Environment Oriented Numerical Modeling*. Ed. by Nakajima, H. Rijeka: IntechOpen. Chap. 17, pp. 459–487.

Aström, K. J. and Murray, R. M. (2009). *Feedback Systems An Introduction for Scientists and Engineers*.

Azmir, J., Zaidul, I. S. M., Rahman, M. M., Sharif, K. M., Mohamed, A., Sahena, F., Jahurul, M. H. A., Ghafoor, K., Norulaini, N. A. N., and Omar, A. K. M. (2013). "Techniques for extraction of bioactive compounds from plant materials : A review". In: *Journal of Food Engineering* 117.4, pp. 426–436.

Ballabio, D. and Todeschini, R. (2009). "Multivariate Classification for Qualitative Analysis". In: *Infrared Spectroscopy for Food Quality Analysis and Control*. Ed. by Sun, D.-W. New York, NY: Academic Press. Chap. 4, pp. 83–104.

Bart, H.-J. and Pilz, S. (2011). *Industrial Scale Natural Products Extraction*. Weinheim, Germany All: Wiley-VCH Verlag & Co. KGaA, pp. 1–291.

Bas, D. and Boyaci, I. H. (2007). "Modeling and optimization I: Usability of response surface methodology". In: *Journal of Food Engineering* 78.3, pp. 836–845.

Baum, A., Dominiak, M., Vidal-Melgosa, S., Willats, W. G. T., Søndergaard, K. M., Hansen, P. W., Meyer, A. S., and Mikkelsen, J. D. (Sept. 2017). "Prediction of Pectin Yield and Quality by FTIR and Carbohydrate Microarray Analysis". In: *Food and Bioprocess Technology* 10.1, pp. 143–154.

Bee Lin, C. and Yek Cze, C. (2018). "Drying Kinetics and Optimisation of Pectin Extraction from Banana Peels via Response Surface Methodology". In: *MATEC Web of Conferences* 152.

Bergeron, C., Carrier, D. J., and Ramaswamy, S. (2012). *Biorefinery Co-Products*. Chichester, UK: John Wiley & Sons, Ltd Registered, pp. 1–351.

Berget, I. and Næs, T. (Mar. 2002a). "Optimal Sorting of Raw Materials, Based on the Predicted End-Product Quality". In: *Quality Engineering* 14.3, pp. 459–478.

— (June 2002b). "Sorting of raw materials with focus on multiple end-product properties". In: *Journal of Chemometrics* 16.6, pp. 263–273.

Bevilacqua, M., Bucci, R., Magri, A. D., Magri, A. L., Nescatelli, R., and Marini, F. (2013). "Classification and Class-Modelling". In: *Data Handling in Science and Technology*. Ed. by Marini, F. Vol. 28. Elsevier B.V. Chap. 5, pp. 171–233.

Bichsel, B., Gal, S., and Signer, R. (1976). "Diffusion phenomena during the decafeination of coffee beans". In: *Journal of Food Technology* 11, pp. 637–646.

Biechele, P., Busse, C., Solle, D., Scheper, T., and Reardon, K. (2015). *Sensor systems for bioprocess monitoring*.

Bogaerts, P. and Wouwer, A. V. (2004). "Parameter identification for state estimation - Application to bioprocess software sensors". In: *Chemical Engineering Science* 59.12, pp. 2465–2476.

Bonvin, D., Srinivasan, B., and Ruppen, D. (2001). "Dynamic Optimization in the Batch Chemical Industry". In: *Chem. Process Control – CPC VI* AIChE Symp.97, pp. 255–273.

Both, S., Eggersglüß, J., Lehnberger, A., Schulz, T., Schulze, T., and Strube, J. (2013). "Optimizing established processes like sugar extraction from sugar beets - Design of experiments versus physicochemical modeling". In: *Chemical Engineering and Technology* 36.12, pp. 2125–2136.

Both, S., Koudous, I., Jenelten, U., and Strube, J. (2014). "Model-based equipment-design for plant-based extraction processes – considering botanic and thermodynamic aspects". In: *Comptes Rendus Chimie* 17.3, pp. 187–196.

Box, G. E. P. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley, p. 669.

Brereton, R. G. and Lloyd, G. R. (Apr. 2014). "Partial least squares discriminant analysis: taking the magic away". In: *Journal of Chemometrics* 28.4, pp. 213–225.

Bro, R. and Smilde, A. K. (2014). "Principal component analysis". In: *Analytical Methods* 6.9, pp. 2812–2831.

Brun, R., Kühni, M., Siegrist, H., Gujer, W., and Reichert, P. (2002). "Practical identifiability of ASM2d parameters - Systematic selection and tuning of parameter subsets". In: *Water Research* 36.16, pp. 4113–4127.

Brun, R., Reichert, P., and Künsch, H. R. (Apr. 2001). "Practical identifiability analysis of large environmental simulation models". In: *Water Resources Research* 37.4, pp. 1015–1030.

Brüniche-Olsen, H. (1962). *Solid-liquid extraction : With particular reference to extraction of sugar from sugar-beets*. eng. Nyt Nordisk Forlag, 463 s.

Buci-Koji, A., Planini, M., Tomas, S., Bili, M., and Veli, D. (2006). "Study of solid–liquid extraction kinetics of total polyphenols from grape seeds". In:

Cacace, J. and Mazza, G. (Oct. 2003). "Mass transfer process during extraction of phenolic compounds from milled berries". In: *Journal of Food Engineering* 59.4, pp. 379–389.

Canettieri, E. V., Rocha, G. J. d. M., Carvalho, J. A. de, and de Almeida e Silva, J. B. (2007). "Optimization of acid hydrolysis from the hemicellulosic fraction of Eucalyptus grandis residue using response surface methodology". In: *Bioresource Technology* 98.2, pp. 422–428.

Canteri-Schemin, M., Fertonani, H., Waszczynskyj, N., and Wosiacki, G. (2005). "Extraction of pectin from apple pomace". In: *Brazilian Archives of Biology and Technology* 48.2, pp. 259–266.

Chan, C.-H., Yusoff, R., and Ngoh, G.-C. (2014). "Modeling and kinetics study of conventional and assisted batch solvent extraction". In: *Chemical Engineering Research and Design* 92.6, pp. 1169–1186.

Chauchan, A., Mittu, B., and Chauchan, P. (2015). "Analytical Method Development and Validation: A Concise Review". In: *Journal of Analytical & Bioanalytical Techniques* 6.1, pp. 1–5.

Chemat, F. and Cravotto, G. (2013). *Microwave-assisted Extraction for Bioactive Compounds*. New York, NY: Springer, p. 248.

Chiang, D., Lin, C. Y., Hu, C. T., and Lee, S. (2018). "Caffeine Extraction from Raw and Roasted Coffee Beans". In: *Journal of Food Science* 83.4, pp. 975–983.

Cho, Y. J. and Hwang, J. K. (2000). "Modeling the yield and intrinsic viscosity of pectin in acidic solubilization of apple pomace". In: *Journal of Food Engineering* 44.2, pp. 85–89.

Chua, B. L., Ng, Y. K., and Ali, A. (2018). "Ultrasound Assisted Extraction of Pectin from Dragon Fruit Peels". In: *Journal of Engineering Science and Technology* 13.Special Issue July, pp. 65–81.

Ciriminna, R., Chavarría-Hernández, N., Inés Rodríguez Hernández, A., and Pagliaro, M. (July 2015). "Pectin: A new perspective from the biorefinery standpoint". In: *Biofuels, Bioproducts and Biorefining* 9.4, pp. 368–377.

Ciriminna, R., Fidalgo, A., Delisi, R., Ilharco, I. M., and Pagliaro, M. (2016). "Pectin Production and Global Market". In: *Agro Food Industry Hi Tech* 27(5), pp. 17–20.

CP Kelco (2010). *GENU® pectin Book*. Tech. rep., pp. 1–36.

Crank, J. (1975). *The mathematics of diffusion*. London, UK: Oxford University Press, pp. 1–414.

Deshpande, A. A., Ramya, A., Vishweshwar, V., Deshpande, G. R., and Roy, A. K. (Dec. 2014). "Applications of Gage Reproducibility & Repeatability (GRR): Understanding and Quantifying the Effect of Variations from Different Sources on a Robust Process Development". In: *Organic Process Research & Development* 18.12, pp. 1614–1621.

Dewasme, L., Goffaux, G., Hantson, A. L., and Wouwer, A. V. (2013). "Experimental validation of an Extended Kalman Filter estimating acetate concentration in E. coli cultures". In: *Journal of Process Control* 23.2, pp. 148–157.

DiFoggio, R. (1995). "Examination of Some Misconceptions About Near-Infrared Analysis". In: *Applied Spectroscopy* 49.1, pp. 67–75.

Dochain, D., ed. (Jan. 2008). *Bioprocess Control*. London, UK: ISTE.

Dochain, D. and Vanrolleghem, P. (2010). "Identification of Bioprocess Models". In: *Bioprocess Control*, pp. 47–78.

Dranca, F. and Oroian, M. (2018). “Extraction, purification and characterization of pectin from alternative sources with potential technological applications, Extraction, purification and characterization of pectin from alternative sources with potential technological applications”. und. In: *Food Research International* 113, pp. 327–350.

Durán, R., Villa, A. L., Ribeiro, R., and Rabi, J. A. (Jan. 2015). “Pectin Extraction from Mango Peels in Batch Reactor: Dynamic One-Dimensional Modeling and Lattice Boltzmann Simulation”. In: *Chemical Product and Process Modeling* 10.3, pp. 203–210.

Durand, A. (2003). “Bioreactor designs for solid state fermentation”. In: *Biochemical Engineering Journal* 13, pp. 113–125.

Engelsen, S. B., Mikkelsen, E., and Munck, L. (1998). “New approaches to rapid spectroscopic evaluation of properties in pectic polymers”. In: *Progress in Colloid and Polymer Science* 108, pp. 166–174.

European Commission (2012). *Innovating for sustainable growth: A bioeconomy for Europe*. Tech. rep., pp. 1–64.

— (2017). *Commission Expert Group on Bio-based Products Final Report*. Tech. rep. November, pp. 1–60.

— (2018). *A sustainable Bioeconomy for Europe : strengthening the connection between economy, society and the environment*. Tech. rep., pp. 1–107.

Food and Drug Administration (2004). *PAT Guidance for Industry — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. Tech. rep., p. 16.

Gan, C.-Y. and Latiff, A. A. (Jan. 2011). “Extraction of antioxidant pectic-polysaccharide from mangosteen (Garcinia mangostana) rind: Optimization using response surface methodology”. In: *Carbohydrate Polymers* 83.2, pp. 600–607.

Geladi, P. and Kowalski, B. R. (1986). “Partial least-squares regression: a tutorial”. In: *Analytica Chimica Acta* 185, pp. 1–17.

Gemperline, P. J., Webber, L. D., and Cox, F. O. (1989). “Raw Materials Testing Using Soft Independent Modeling of Class Analogy Analysis of Near-Infrared Reflectance Spectra”. In: *Analytical Chemistry* 61.2, pp. 138–144.

Gernaey, K. V., Lantz, A. E., Tufvesson, P., Woodley, J. M., and Sin, G. (2010). “Application of mechanistic models to fermentation and biocatalysis for next-generation processes”. In: *Trends in Biotechnology* 28.7, pp. 346–354.

Gernaey, K. V., Woodley, J., Lantz, A. E., and Sin, G. (Sept. 2009). “Mechanistic models and advanced model analysis within a PAT framework”. In: *New Biotechnology* 25.Suppl. 1, S242.

Girisuta, B., Janssen, L. P., and Heeres, H. J. (2007). “Kinetic study on the acid-catalyzed hydrolysis of cellulose to levulinic acid”. In: *Industrial and Engineering Chemistry Research* 46.6, pp. 1696–1708.

Gorissen, B. L., Yanıkoğlu, İ., and Hertog, D. den (June 2015). “A practical guide to robust optimization”. In: *Omega* 53, pp. 124–137.

Gower, J. (1995). "A general theory of biplots". In: *Recent Advances in Descriptive Multivariate Statistics*. Ed. by Krzanowski, W. Oxford University Press, pp. 283–303.

Gray, N. (2014). *Pectin prices: Tight supply and high raw material costs lead to price rise for Andre Pectin*. URL: `https://www.foodnavigator.com/Article/2014/08/05/Pectin-prices-Tight-supply-and-high-raw-material-costs-lead-to-price-rise-for-Andre-Pectin` (visited on July 6, 2018).

Green, J. M. (1996). "A Practical Guide to Analytical Method Validation". In: *Analytical Chemistry* 68, 305A–309A.

Grewal, M. S. and Andrews, A. P. (2008). *Kalman filtering - theory and practice using MATLAB*. Third. Hoboken, New Jersey: John Wiley & Sons, pp. 1–575.

Gudi, G., Krähmer, A., Koudous, I., Strube, J., and Schulz, H. (2015). "Infrared and Raman spectroscopic methods for characterization of Taxus baccata L. - Improved taxane isolation by accelerated quality control and process surveillance". In: *Talanta* 143, pp. 42–49.

Hangos, K. M., Bokor, J., and Szederkényi, G. (2004). *Analysis and Control of Nonlinear Process Systems*.

Helton, J. C. and Davis, F. J. (2003). "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems". In: *Reliability Engineering & System Safety* 81, pp. 23–69.

Hermann, R. and Krener, A. J. (1977). "Nonlinear Controllability and Observability". In: *IEEE Transactions on Automatic Control* 22.5, pp. 728–740.

Herodež, Š. S., Hadolin, M., Škerget, M., and Knez, Ž. (2003). "Solvent extraction study of antioxidants from Balm (Melissa officinalis L.) leaves". In: *Food Chemistry* 80.2, pp. 275–282.

Hojnik, M., Škerget, M., and Knez, Ž. (2008). "Extraction of lutein from Marigold flower petals - Experimental kinetics and modelling". In: *LWT - Food Science and Technology* 41.10, pp. 2008–2016.

ICH Expert Working Group (2000). *Q7 - Good Manufacturing Practice Guide for Active Pharmaceutical Ingredients*. Tech. rep. November, p. 49.

— (2005). *Q2 (R1) - Validation of Analytical Procedures : Text and Methodology*. Tech. rep., p. 17.

— (2009). *Guideline Pharmaceutical Development Q8(R2)*. Tech. rep., pp. 1–24.

Ji, J.-b., Lu, X.-h., Cai, M.-q., and Xu, Z.-c. (2006). "Improvement of leaching process of Geniposide with ultrasound". eng. In: *Ultrasonics Sonochemistry* 13.5, pp. 455–462.

Joana Gil-Chávez, G., Villa, J. A., Fernando Ayala-Zavala, J., Basilio Heredia, J., Sepulveda, D., Yahia, E. M., and González-Aguilar, G. A. (2013). "Technologies for Extraction and Production of Bioactive Compounds to be Used as Nutraceuticals and Food Ingredients: An Overview". In: *Comprehensive Reviews in Food Science and Food Safety* 12.1, pp. 5–23. arXiv: `arXiv:1011.1669v3`.

Jokic, S., Velic, D., Bilic, M., Ana Bucic-Kojic, Planinic, M., and Tomasa, S. (2010). "Modelling of the process of solid-liquid extraction of total polyphenols from soybeans". In: *Czech Journal of Food Sciences* 28.3, pp. 206–212.

Jørgensen, K., Mevik, B., and Næs, T. (2007). "Combining designed experiments with several blocks of spectroscopic data". In: *Chemometrics and Intelligent Laboratory Systems* 88.2, pp. 154–166.

Jørgensen, K. and Næs, T. (Feb. 2004). "A design and analysis strategy for situations with uncontrolled raw material variation". In: *Journal of Chemometrics* 18.2, pp. 45–52.

Jose, G. E., Folque, F., Menezes, J. C., Werz, S., Strauss, U., and Hakemeyer, C. (2011). "Predicting mAb product yields from cultivation media components, using near-infrared and 2D-fluorescence spectroscopies". In: *Biotechnology Progress* 27.5, pp. 1339–1346.

Jothi, D. (2008). "Extraction of natural dyes from African marigold flower (Tagetes ereectal) for textile coloration". In: *Autex Research Journal* 8.2, pp. 49–53.

Kalman, R. (1960). "On the general theory of control systems". In: *IRE Transactions on Automatic Control* 4.3, pp. 110–110.

Kalman, R. E. (1960). "Contributions to the theory of optimal control". In: *Boletin de la Sociedad Matematica Mexicana* 5.2, pp. 102–119.

Kassing, M., Jenelten, U., Schenk, J., Hänsch, R., and Strube, J. (2012). "Combination of rigorous and statistical modeling for process development of plant-based extractions based on mass balances and biological aspects". In: *Chemical Engineering and Technology* 35.1, pp. 109–132.

Kassing, M., Jenelten, U., Schenk, J., and Strube, J. (Mar. 2010a). "A New Approach for Process Development of Plant-Based Extraction Processes". In: *Chemical Engineering & Technology* 33.3, pp. 377–387.

Kassing, M., Jenelten, U., Schenk, J., and Strube, J. (2010b). "A new approach for process development of plant-based extraction processes". In: *Chemical Engineering and Technology* 33.3, pp. 377–387.

Kaya, M., Sousa, A. G., Crepeau, M.-J., Sorensen, S. O., and Ralet, M.-C. (Oct. 2014). "Characterization of citrus pectin samples extracted under different conditions: influence of acid type and pH of extraction". In: *Annals of Botany* 114.6, pp. 1319–1326.

Kennard, R. W. and Stone, L. A. (Feb. 1969). "Computer Aided Design of Experiments". In: *Technometrics* 11.1, pp. 137–148.

Khan, M. K., Abert-Vian, M., Fabiano-Tixier, A.-S., Dangles, O., and Chemat, F. (2010). "Ultrasound-assisted extraction of polyphenols (flavanone glycosides) from orange (Citrus sinensis L.) peel". In: *Food Chemistry* 119.2, pp. 851–858.

Kirdar, A. O., Chen, G., Weidner, J., and Rathore, A. S. (2009). "Application of near-infrared (NIR) spectroscopy for screening of raw materials used in the cell culture medium for the production of a recombinant therapeutic protein". In: *Biotechnology Progress* 26.2, pp. 527–531.

Kliemann, E., De Simas, K., Amante, E., Prudêncio, E., Teófilo, R., Ferreira, M., and Amboni, R. (2009). "Optimisation of pectin acid extraction from passion fruit peel (Passiflora edulis flavicarpa) using response surface methodology". In: *International Journal of Food Science and Technology* 44.3, pp. 476–483.

Krishna, C. (2008). “Solid-State Fermentation Systems — An Overview Solid-State Fermentation Systems — An Overview”. In: *Critical Reviews in Biotechnology,* 25.1-2, pp. 1–30.

Ku, C. and Mun, S. (2008). “Optimization of the extraction of anthocyanin from Bokbunja (Rubus coreanus Miq.) marc produced during traditional wine processing and characterization of the extracts”. In: *Bioresource Technology* 99.17, pp. 8325–8330.

Kulkarni, S. and Vijayanand, P. (2010). “Effect of extraction conditions on the quality characteristics of pectin from passion fruit peel (Passiflora edulis f. flavicarpa L.)” In: *LWT - Food Science and Technology* 43.7, pp. 1026–1031.

Lanan, M. (2008). “QbD for Raw Materials”. In: *Quality by Design for Biopharmaceuticals*. Ed. by Rathore, A. S. and Mhatre, R. Hoboken, NJ, USA: John Wiley & Sons, Inc. Chap. 11, pp. 193–209.

Langhurst, L. F. (1951). “Solvent Extraction Processes”. In: *Soybeans and soybean products*. Ed. by Markley, K. S. New York, NY: Interscience Publishers. inc. Chap. XV, pp. 541–581.

Le Clef, E. and Kemper, T. (2015). “Sunflower Seed Preparation and Oil Extraction”. In: *Sunflower: Chemistry, Production, Processing, and Utilization*. AOCS Press. Chap. 8, pp. 187–226.

Lebreton-Decoster, C., Rousselle, P., Laperdrix, C., Lubrano, C., Robin, J. R., and Coulomb, B. (2011). “Oligogalacturonides improve tissue organization of in vitro reconstructed skin”. In: *International Journal of Cosmetic Science* 33.5, pp. 455–461.

Lee, S., Choi, H., Cha, K., and Chung, H. (2013). “Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha”. In: *Microchemical Journal* 110, pp. 739–748.

Lencastre Fernandes, R., Bodla, V. K., Carlquist, M., Heins, A.-L., Eliasson Lantz, A., Sin, G., and Gernaey, K. V. (2013). “Applying Mechanistic Models in Bioprocess Development”. In: *Measurement, Monitoring, Modelling and Control of Bioprocesses*. Springer-Verla Berlin, pp. 137–166.

Li Vigni, M., Durante, C., and Cocchi, M. (2013). “Exploratory Data Analysis”. In: *Data Handling in Science and Technology*. Ed. by Marini, F. 1st ed. Vol. 28. Elsevier B.V. Chap. 3, pp. 55–126.

Lopez da Silva, J. A. and Rao, M. A. (2006). “Pectins: Structure, Functionality, and Uses”. In: *Food Polysaccharides and Their Applications*. Ed. by Stephen, A., Phillips, G., and Williams, P. Second. CRC Press-Taylor & Francis Group. Chap. 11, pp. 353–411.

Lopez, T. (1999). “Robust Dynamic State Estimation of Nonlinear Plants”. In: *AlChE Journal* 45.1.

Luenberger, D. (1971). “An introduction to observers”. In: *IEEE Transactions on automatic control* 16.6, pp. 596–602.

Lupi, F. R., Gabriele, D., Seta, L., Baldino, N., Cindio, B. de, and Marino, R. (2015). "Rheological investigation of pectin-based emulsion gels for pharmaceutical and cosmetic uses". In: *Rheologica Acta* 54.1, pp. 41–52.

Luttmann, R., Bracewell, D. G., Cornelissen, G., Gernaey, K. V., Glassey, J., Hass, V. C., Kaiser, C., Preusse, C., Striedner, G., and Mandenius, C.-F. (Aug. 2012). "Soft sensors in bioprocessing: A status report and recommendations". In: *Biotechnology Journal* 7.8, pp. 1040–1048.

Mandenius, C.-F. and Titchener-Hooker, N. J., eds. (2013). *Measurement, Monitoring, Modelling and Control of Bioprocesses*. Vol. 132. Advances in Biochemical Engineering/Biotechnology. Berlin, Heidelberg: Springer Berlin Heidelberg.

Marathe, S. J., Jadhav, S. B., Bankar, S. B., and Singhal, R. S. (2017). "Enzyme-Assisted Extraction of Bioactives". In: *Food Bioactives: Extraction and Biotechnology Applications*. Ed. by Puri, M. Cham: Springer International Publishing, pp. 171–201.

Maric, M., Grassino, A. N., Zhu, Z., Barba, F. J., Brncic, M., and Brncic, S. R. (2018). "An overview of the traditional and innovative approaches for pectin extraction from plant food wastes and by-products: Ultrasound-, microwaves-, and enzyme-assisted extraction". eng. In: *Trends in Food Science and Technology* 76, pp. 28–37.

Masmoudi, M., Besbes, S., Chaabouni, M., Robert, C., Paquot, M., Blecker, C., and Attia, H. (2008). "Optimization of pectin extraction from lemon by - product with acidified date juice using response surface methodology". In: *Carbohydr Polym.* 74, pp. 185–192.

Massart, D., Vandeginste, B., Deming, S., Michotte, Y., and Kaufman, L. (1988). *Chemometrics: a textbook*. Elsevier Science B.V., p. 488.

Maxwell, E. G., Belshaw, N. J., Waldron, K. W., and Morris, V. J. (2012). "Pectin: An emerging new bioactive food polysaccharide". In: *Trends in Food Science & Technology* 24.2, pp. 64–73.

May, C. D. (1997). "Pectins". In: *Thickening and Gelling Agents for Food*. Ed. by Imeson, A. P. Second. Chapman & Hall, pp. 230–261.

May, C. D. (1990). "Industrial Pectins: Sources, Production and Applications". In: *Carbohydrate Polymers* 12, pp. 79–99.

McHugh, D. J. (2003). *A Guide to the Seaweed Industry*. 441. Rome: Food and Agriculture Organization of the United Nations, p. 105.

Meireles, M. A. a. (2009). "1 Extraction and Purification of Bioactive Compounds". In: pp. 1–8.

Mevik, B., Segtnan, V. H., and Næs, T. (2004). "Ensemble methods and partial least squares regression". In: *Journal of Chemometrics* 18.11, pp. 498–507.

Minjares-Fuentes, R., Femenia, A., Garau, M., Meza-Velázquez, J., Simal, S., and Rosselló, C. (2014). "Ultrasound-assisted extraction of pectins from grape pomace using citric acid: A response surface methodology approach". In: *Carbohydrate Polymers* 106.1, pp. 179–189.

Minkov, S., Minchev, A., and Paev, K. (1996). "Modelling of the hydrolysis and extraction of apple pectin". In: *Journal of Food Engineering* 29.1, pp. 107–113.

Mohd Ali, J., Ha Hoang, N., Hussain, M., and Dochain, D. (May 2015). "Review and classification of recent observers applied in chemical process systems". In: *Computers & Chemical Engineering* 76, pp. 27–41.

Murtagh, F. and Legendre, P. (2014). "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" In: *Journal of Classification* 31, pp. 274–295.

Mussatto, S. I. and Teixeira, J. A. (2010). "Lignocellulose as raw material in fermentation processes". In: *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology*. Ed. by Méndez-Vilas, A. Badajoz, Spain: Formatex Research Center, pp. 897–907.

Næs, T., Isaksson, T., Fearn, T., and Davies, T. (2004). *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester, UK: NIR Publications, p. 344.

Nawar, S. and Mouazen, A. M. (2017). "Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon". In: *Sensors* 17.10, pp. 1–22.

Newman, D. J. and Cragg, G. M. (2016). "Natural Products as Sources of New Drugs from 1981 to 2014". In: *Journal of Natural Products* 79, pp. 629–661.

Newman, D. J., Cragg, M., Snader, K. M., and December, C. (2000). "The influence of natural products upon drug discovery". In: *Natural Product Reports* 17, pp. 215–234.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., and Engelsen, S. B. (Mar. 2000). "Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy". In: *Applied Spectroscopy* 54.3, pp. 413–419.

Oliveira, T., Rosa, M., Cavalcante, F., Pereira, P., Moates, G., Wellner, N., Mazzetto, S., Waldron, K., and Azeredo, H. (2016). "Optimization of pectin extraction from banana peels with citric acid by using response surface methodology". In: *Food Chemistry* 198, pp. 113–118.

Omlin, M. and Reichert, P. (1999). "A comparison of techniques for the estimation of model prediction uncertainty". In: *Ecological Modelling* 115.1, pp. 45–59.

Osburn, J. and Katz, D. (1944). "Structure as a variable in the application of diffusion theory to extraction". eng. In: *Transactions of the American Institute of Chemical Engineers* 40.5, pp. 0511–0531.

Pagán, J. and Ibarz, A. (Feb. 1999). "Extraction and rheological properties of pectin from fresh peach pomace". In: *Journal of Food Engineering* 39.2, pp. 193–201.

Panchev, I. N., Kirtchev, N. A., and Kratchanov, C. (1989). "Kinetic Model of Pectin Extraction". In: *Carbohydrate Polymers* 11, pp. 193–204.

Pearsall, S. M., Rowley, C. N., and Berry, A. (2015). "Advances in Pathway Engineering for Natural Product Biosynthesis". In: *ChemCatChem* 7, pp. 3078–3093.

Peleg, M. (1988). "An Empirical Model for the Description of Moisture Sorption Curves". eng. In: *Journal of Food Science* 53.4, pp. 1216–1217.

Pinelo, M., Sineiro, J., and Nú, M. J. (2005). "Mass transfer during continuous solid–liquid extraction of antioxidants from grape byproducts". In: *Journal of Food Engineering* 77, pp. 57–63.

Pingret, D., Fabiano-Tixier, A.-S., Bourvellec, C. L., Renard, C. M., and Chemat, F. (2012). "Lab and pilot-scale ultrasound-assisted water extraction of polyphenols from apple pomace". In: *Journal of Food Engineering* 111.1, pp. 73–81.

Pinheiro, E., Silva, I., Gonzaga, L., Amante, E., Teófilo, R., Ferreira, M., and Amboni, R. (2008). "Optimization of extraction of high-ester pectin from passion fruit peel (Passiflora edulis flavicarpa) with citric acid by using response surface methodology". In: *Bioresource Technology* 99.13, pp. 5561–5566.

Prakash Maran, J., Manikandan, S., Thirugnanasambandham, K., Vigna Nivetha, C., and Dinesh, R. (2013a). "Box-Behnken design based statistical modeling for ultrasound-assisted extraction of corn silk polysaccharide". In: *Carbohydrate Polymers* 92.1, pp. 604–611.

Prakash Maran, J., Sivakumar, V., Thirugnanasambandham, K., and Sridhar, R. (2013b). "Optimization of microwave assisted extraction of pectin from orange peel". In: *Carbohydrate Polymers* 97.2, pp. 703–709.

Prunescu, R. M. and Sin, G. (2013). "Dynamic modeling and validation of a lignocellulosic enzymatic hydrolysis process – A demonstration scale study". In: *Bioresource Technology* 150, pp. 393–403.

Puri, M., Sharma, D., and Barrow, C. J. (2012). "Enzyme-assisted extraction of bioactives from plants". In: *Trends in Biotechnology* 30.1, pp. 37–44.

Putnik, P., Bursać Kovačević, D., Režek Jambrak, A., Barba, F. J., Cravotto, G., Binello, A., Lorenzo, J. M., and Shpigelman, A. (2017). "Innovative "green" and novel strategies for the extraction of bioactive added value compounds from citruswastes - A review". In: *Molecules* 22.5.

Quanhong, L. and Caili, F. (2005). "Application of response surface methodology for extraction optimization of germinant pumpkin seeds protein". In: *Food Chemistry* 92.4, pp. 701–706.

Randek, J. and Mandenius, C. F. (2017). "On-line soft sensing in upstream bioprocessing". In: *Critical Reviews in Biotechnology*, pp. 1–16.

Rathore, A. S., Bhambure, R., and Ghare, V. (2010). "Process Analytical Technology (PAT) for Biopharmaceutical Products: Part I. Concepts and Applications". In: *Biotechnology and Bioengineering* 105.2, pp. 276–284.

Rathore, A. S., Bansal, A., and Hans, J. (2013). "Knowledge Management and Process Monitoring of Pharmaceutical Processes in the Quality by Design Paradigm". In: *Measurement, Monitoring, Modelling and Control of Bioprocesses*. Ed. by Mandenius, C.-F. and Titchener-Hooker, N. J. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 217–247.

Rathore, A. S., Mittal, S., Pathak, M., and Arora, A. (2014). "Guidance for performing multivariate data analysis of bioprocessing data: Pitfalls and recommendations". In: *Biotechnology Progress* 30.4, pp. 967–973.

Rathore, A. S. and Winkle, H. (Jan. 2009). "Quality by design for biopharmaceuticals". In: *Nature Biotechnology* 27.1, pp. 26–34.

Reverchon, E. and De Marco, I. (Sept. 2006). "Supercritical fluid extraction and fractionation of natural matter". In: *The Journal of Supercritical Fluids* 38.2, pp. 146–166.

Rhein-Knudsen, N., Ale, M. T., and Meyer, A. S. (2015). "Seaweed hydrocolloid production: An update on enzyme assisted extraction and modification technologies". In: *Marine Drugs* 13.6, pp. 3340–3359.

Rinnan, Å., Berg, F. van den, and Engelsen, S. B. (2009). "Review of the most common pre-processing techniques for near-infrared spectra". In: *TrAC Trends in Analytical Chemistry* 28.10, pp. 1201–1222.

Rombaut, N., Tixier, A.-S., Bily, A., and Chemat, F. (2014). "Green extraction processes of natural products as tools for biorefinery". In: *Biofuels, Bioprod. Bioref.* 8, pp. 530–544.

Rostami, M., Farzaneh, V., Boujmehrani, A., Mohammadi, M., and Bakhshabadi, H. (2014). "Optimizing the extraction process of sesame seed's oil using response surface method on the industrial scale". In: *Industrial Crops and Products* 58, pp. 160–165.

Sabater, C., Corzo, N., Olano, A., and Montilla, A. (2018). "Enzymatic extraction of pectin from artichoke (Cynara scolymus L.) by-products using Celluclast®1.5L". In: *Carbohydrate Polymers* 190, pp. 43–49.

Sagmeister, P., Wechselberger, P., Jazini, M., Meitz, A., Langemann, T., and Herwig, C. (2013a). "Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development". In: *Chemical Engineering Science* 96, pp. 190–198.

— (2013b). "Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development". In: *Chemical Engineering Science* 96, pp. 190–198.

Sakai, T., Sakamoto, T., Hallaert, J., and Vandamme, E. (1993). "Pectin, Pectinase, and Protopectinase - Production, Properties and Applications". In: *Advances in Applied Microbiology* 39, pp. 213–294.

Sarker, S. D. and Nahar, L. (2012). *Natural Products Isolation: Methods and Protocols*. Springer New York Dordrecht Heidelberg London, pp. 1–557.

Sathish, S., Narendrakumar, G., Sundari, N., Amarnath, M., and Thayyil, P. (2016). "Extraction of pectin from used citrus Limon and optimization of process parameters using response surface methodology". In: *Research Journal of Pharmacy and Technology* 9.12, pp. 1451–1456.

Schneider, R. and Georgakis, C. (2013). "How to NOT make the extended kalman filter fail". In: *Industrial and Engineering Chemistry Research* 52.9, pp. 3354–3362.

Scott-Thomas, C. (2013). *Yantai Andre Pectin increases pectin prices on raw material*. URL: `https://www.foodnavigator.com/Article/2013/10/30/Yantai-Andre-Pectin-increases-pectin-prices-on-raw-material-shortage` (visited on July 6, 2018).

Seader, J. D. and Henley, E. J. (2006). *Separation process principles*. Wiley, p. 756.

Sebaoui, O., Moussaoui, R., Kadi, H., Michaud, P., and Delattre, C. (2017). "Kinetic Modeling of Pectin Extraction from Wasted Citrus Lemon L." In: *Waste and Biomass Valorization* 8.7, pp. 2329–2337.

Seikova, I., Simeonov, E., and Ivanova, E. (2004). "Protein leaching from tomato seed—experimental kinetics and prediction of effective diffusivity". In: *Journal of Food Engineering* 61.2, pp. 165–171.

Seixas, F., Fukuda, D., Turbiani, F., Garcia, P., Petkowicz, C., Jagadevan, S., and Gimenes, M. (2014). "Extraction of pectin from passion fruit peel (Passiflora edulis f.flavicarpa) by microwave-induced heating". In: *Food Hydrocolloids* 38, pp. 189–192.

Silva, E. M., Rogez, H., and Larondelle, Y. (2007). "Optimization of extraction of phenolics from Inga edulis leaves using response surface methodology". In: *Separation and Purification Technology* 55.3, pp. 381–387.

Simeonov, E., Tsibranska, I., and Minchev, A. (1999). "Solid–liquid extraction from plants — experimental kinetics and modelling".

Simeonov, E. and Nedialcova, J. (2006). "Process control and estimation of multistage extraction using characteristic function approach". In: *Journal of the University of Chemical Technology and Metallurgy* 41.4, pp. 445–448.

Simon, L. L., Pataki, H., Marosi, G., Meemken, F., Hungerbühler, K., Baiker, A., Tummala, S., Glennon, B., Kuentz, M., Steele, G., Kramer, H. J. M., Rydzak, J. W., Chen, Z., Morris, J., Kjell, F., Singh, R., Gani, R., Gernaey, K. V., Louhi-Kultanen, M., O'Reilly, J., Sandler, N., Antikainen, O., Yliruusi, J., Frohberg, P., Ulrich, J., Braatz, R. D., Leyssens, T., Stosch, M. von, Oliveira, R., Tan, R. B. H., Wu, H., Khan, M., O'Grady, D., Pandey, A., Westra, R., Delle-Case, E., Pape, D., Angelosante, D., Maret, Y., Steiger, O., Lenner, M., Abbou-Oucherif, K., Nagy, Z. K., Litster, J. D., Kamaraju, V. K., and Chiu, M.-S. (2015). "Assessment of Recent Process Analytical Technology (PAT) Trends: A Multiauthor Review". In: *Organic Process Research & Development* 19.1, pp. 3–62.

Sin, G. and Gernaey, K. (2016). "Data Handling and Parameter Estimation". In: *Experimental Methods in Wastewater Treatment*. IWA Publishing, pp. 201–234.

Sin, G., Meyer, A. S., and Gernaey, K. V. (2010). "Assessing reliability of cellulose hydrolysis models to support biofuel process design-Identifiability and uncertainty analysis". In: *Computers and Chemical Engineering* 34.9, pp. 1385–1392.

Sin, G. and Vanrolleghem, P. A. (2007). "Extensions to modeling aerobic carbon degradation using combined respirometric-titrimetric measurements in view of activated sludge model calibration". In: *Water Research* 41.15, pp. 3345–3358.

Sinclair, W. B. and Eny, D. M. (Mar. 1947). "Ether-Soluble Organic Acids and Buffer Properties of Citrus Peels". In: *Botanical Gazette* 108.3, pp. 398–407.

Sivakumar, V., Vijaeeswarri, J., and Anna, J. L. (2011). "Effective natural dye extraction from different plant materials using ultrasound". In: *Industrial Crops and Products* 33.1, pp. 116–122.

Sixt, M., Gudi, G., Schulz, H., and Strube, J. (2018a). "In-line Raman spectroscopy and advanced process control for the extraction of anethole and fenchone from fennel". In: *Comptes rendus - Chimie* 21.2, pp. 97–103.

Sixt, M., Uhlenbrock, L., and Strube, J. (2018b). "Toward a Distinct and Quantitative Validation Method for Predictive Process Modelling — On the Example of Solid-

Liquid Extraction Processes of Complex Plant Extracts". In: *Processes* 6.66, pp. 1–27.

Skibsted, E. (2011). "Process Analytical Technology Applied to Raw Materials". In: *PAT Applied in Biopharmaceutical Process Development And Manufacturing An Enabling Tool for Quality-by-Design*. Ed. by Undey, C., Low, D., Menezes, J. C., and Koch, M. CRC Press-Taylor & Francis Group. Chap. 7, pp. 127–141.

Skov, T., Honoré, A. H., Jensen, H. M., Næs, T., and Engelsen, S. B. (Sept. 2014). "Chemometrics in foodomics: Handling data structures from multiple analytical platforms". In: *TrAC Trends in Analytical Chemistry* 60, pp. 71–79.

Sonnleitner, B. (2012). "Automated Measurement and Monitoring of Bioprocesses: Key Elements of the M3C Strategy". In: pp. 1–33.

Sorsamäki, L., Kaijaluoto, S., Aaltonen, O., and Nakari-Setälä, T. (2010). "Bark-Based Biorefinery – From Pilot Experiments to Process Model". In: *AIChE Annual Meeting*, pp. 1–14.

Sriamornsak, P. (2003). "Chemistry of Pectin and Its Pharmaceutical Uses : A Review". In: *Silpakorn University International Journal* 3.1-2, pp. 206–228.

Staffas, L., Gustavsson, M., and Mccormick, K. (2013). "Strategies and Policies for the Bioeconomy and Bio-Based Economy: An Analysis of Official National Approaches". In: *Sustainability* 5, pp. 2751–2769.

Sun, D.-W. (2009). *Infrared spectroscopy for food quality analysis and control*. Academic Press, p. 424.

Tiwari, G. and Tiwari, R. (Oct. 2010). "Bioanalytical method validation: An updated review." In: *Pharmaceutical methods* 1.1, pp. 25–38.

Tsibranska, L., Tylkowski, B., Kochanov, R., and Alipieva, K. (2011). "Extraction of biologically active compounds from Sideritis ssp L". eng. In: *Food and Bioproducts Processing* 89.C4, pp. 273–280.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co, p. 688.

Udugama, I. A., Mansouri, S. S., Mitic, A., Flores-Alsina, X., and Gernaey, K. V. (2017). "Perspectives on Resource Recovery from Bio-Based Production Processes : From Concept to Implementation". In: *Processes* 5.48, pp. 1–25.

Uhlenbrock, L., Sixt, M., Tegtmeier, M., Schulz, H., Hagels, H., Ditz, R., and Strube, J. (2018). "Natural Products Extraction of the Future—Sustainable Manufacturing Solutions for Societal Needs". In: *Processes* 6.10.

Villaverde, A. F. and Banga, J. R. (2017). "Structural Properties of Dynamic Systems Biology Models: Identifiability, Reachability, and Initial Conditions". In: *Processes* 5.2, p. 29.

Wang, L. and Weller, C. L. (2006). "Recent advances in extraction of nutraceuticals from plants". In: *Trends in Food Science and Technology* 17.6, pp. 300–312.

Wang, S., Chen, F., Wu, J., Wang, Z., Liao, X., and Hu, X. (2007). "Optimization of pectin extraction assisted by microwave from apple pomace using response surface methodology". In: *Journal of Food Engineering* 78.2, pp. 693–700.

Wang, W., Ma, X., Xu, Y., Cao, Y., Jiang, Z., Ding, T., Ye, X., and Liu, D. (2015). "Ultrasound-assisted heating extraction of pectin from grapefruit peel: Optimiza-

tion and comparison with the conventional method". In: *Food Chemistry* 178, pp. 106–114.

Wang, X. and Lü, X. (2014). "Characterization of pectic polysaccharides extracted from apple pomace by hot-compressed water". In: *Carbohydrate Polymers* 102.1, pp. 174–184.

Widyaningsih, T. D., Widjanarko, S. B., Waziiroh, E., Wijayanti, N., and Maslukhah, Y. L. (2018). "Pilot plant scale extraction of black cincau (Mesona palustris BL) using historical-data response surface methodology". In: *International Food Research Journal* 25.2, pp. 712–719.

Willats, W. G. T., McCartney, L., Mackie, W., and Knox, J. P. (2001). "Pectin: cell biology and prospects for functional analysis". In: *Plant Molecular Biology* 47, pp. 9–27.

Winning, H., Viereck, N., Nørgaard, L., Larsen, J., and Engelsen, S. B. (2007). "Quantification of the degree of blockiness in pectins using H NMR spectroscopy and chemometrics". In: *Food Hydrocolloids* 21, pp. 256–266.

Wongkittipong, R., Prat, L., Damronglerd, S., and Gourdon, C. (2004). "Solid–liquid extraction of andrographolide from plants—experimental study, kinetic reaction and model". In: *Separation and Purification Technology* 40, pp. 147–154.

Woodley, J. M., Breuer, M., and Mink, D. (2013). "Chemical Engineering Research and Design A future perspective on the role of industrial biotechnology for chemicals production". In: *Chemical Engineering Research and Design* 91.10, pp. 2029–2036.

Wu, S. and Chappell, J. (2008). "Metabolic engineering of natural products in plants ; tools of the trade and challenges for the future". In: *Current Opinion in Biotechnology* 19, pp. 145–152.

Xiaobo, Z., Jiewen, Z., Povey, M. J. W., Holmes, M., and Hanpin, M. (2010). "Variables selection methods in near-infrared spectroscopy". In: *Analytica Chimica Acta* 667.1-2, pp. 14–32.

Xu, Y., Zhang, L., Bailina, Y., Ge, Z., Ding, T., Ye, X., and Liu, D. (2014). "Effects of ultrasound and/or heating on the extraction of pectin from grapefruit peel". In: *Journal of Food Engineering* 126, pp. 72–81.

Yu, L. X., Amidon, G., Khan, M. A., Hoag, S. W., Polli, J., Raju, G. K., and Woodcock, J. (July 2014). "Understanding Pharmaceutical Quality by Design". In: *The AAPS Journal* 16.4, pp. 771–783.

Zhang, G., He, L., and Hu, M. (2011). "Optimized ultrasonic-assisted extraction of flavonoids from Prunella vulgaris L. and evaluation of antioxidant activities in vitro". In: *Innovative Food Science and Emerging Technologies* 12.1, pp. 18–25.

Zhang, Y.-Y., Mu, T.-H., and Zhang, M. (2013). "Optimisation of acid extraction of pectin from sweet potato residues by response surface methodology and its antiproliferation effect on cancer cells". In: *International Journal of Food Science and Technology* 48.4, pp. 778–785.

Ziari, H., Ashtiani, F. Z., and Mohtashamy, M. (2010). "Comparing the effectiveness of processing parameters in pectin extraction from apple pomace". In: *Afinidad LXVII* 549, pp. 374–379.

Zykwinska, A., Rondeau-Mouro, C., Garnier, C., Thibault, J.-F., and Ralet, M.-C. (2006). "Alkaline extractability of pectic arabinan and galactan and their mobility in sugar beet and potato cell walls". In: *Carbohydrate Polymers* 65.4, pp. 510–520.

{2.7182818284}