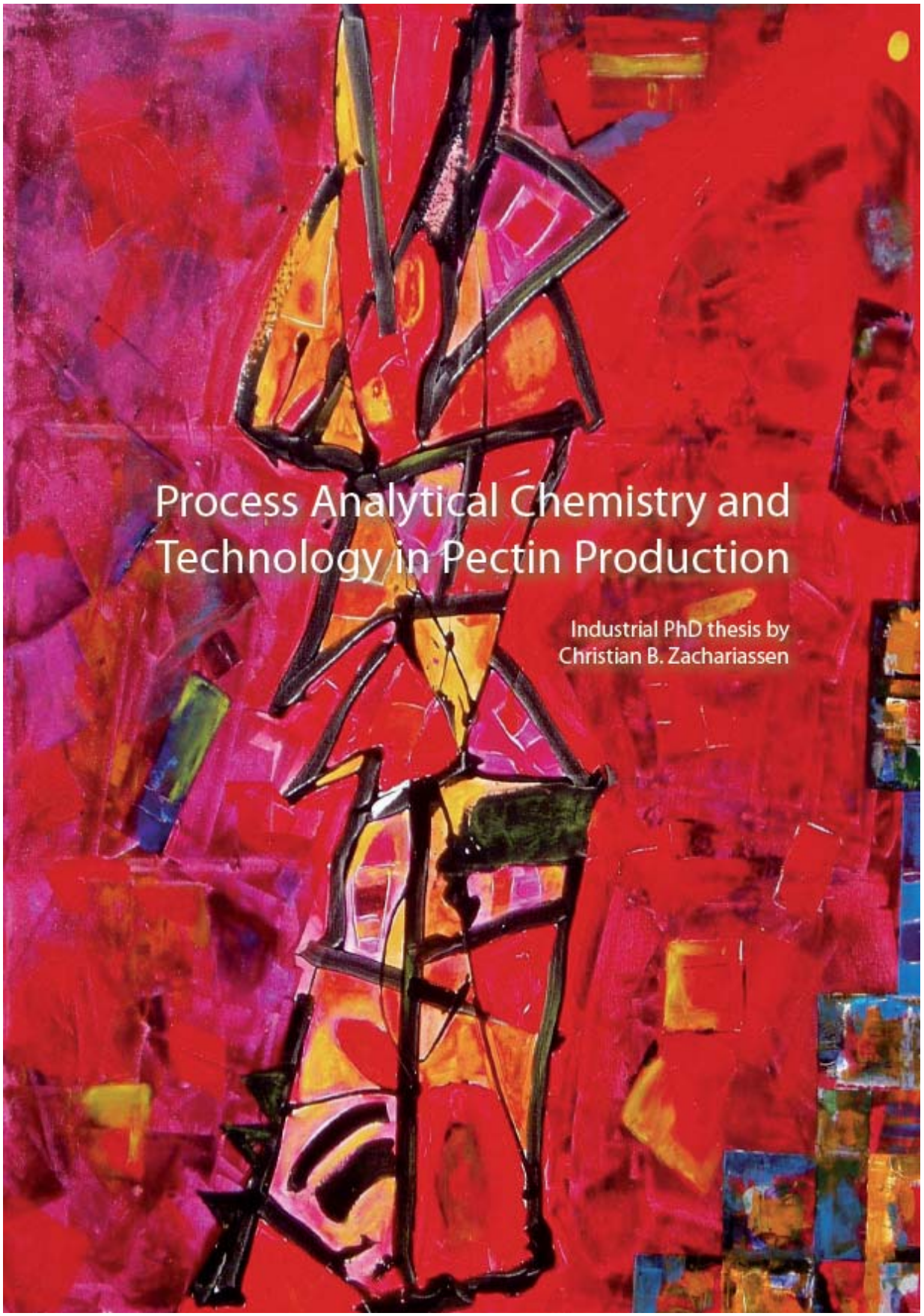

Process Analytical Chemistry and Technology in Pectin Production
Industrial PhD thesis by
Christian B. Zachariassen

FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

Process Analytical Chemistry and Technology in Pectin Production

Industrial PhD thesis by
Christian B. Zachariassen, 2007

Supervisors:
Principal Scientist Jan Larsen[a]
Assistant Professor Frans van den Berg[b]
and Professor Søren Balling Engelsen[b]

[a]CP Kelco, Lille Skensved, Denmark
[b]Quality & Technology, Department of Food Science,
Faculty of Life Sciences, University of Copenhagen, Denmark

No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording without permission from the author.

Cover illustration by Poul-Erik Bomholt Zachariassen
Title: Process Analytical Chemistry and Technology in Pectin Production
© 2007, Christian B. Zachariassen, czachariassen@gmail.com
ISBN: 978-87-7611-183-0
Printed by Samfundslitteratur Grafik, Frederiksberg, Denmark

Preface

This Industrial PhD thesis is written to partly fulfil the requirements for obtaining a PhD degree at University of Copenhagen (KU). The expenses in connection with participation in the Ph.D. project were partly covered by a grant made available by the Ministry of Science, Technology and Innovation. The presented work has been carried out at CP Kelco ApS and the Quality and Technology Group at KU under the supervision of Principal Scientist Jan Larsen (CP Kelco), Assistant Professor Frans van den Berg and Professor Søren Balling Engelsen (KU). The project also included a three month stay at The Chemical and Environmental Research Institute of Barcelona, Spain, visiting Research Professor Romà Tauler and Associate Professor Anna de Juan. Professor Rasmus Bro was there on the sideline co-authoring and pushing things in the right direction. I am very grateful for all your inspiration, guidance, and support. Everything from the smallest details to the greatest lines has been discussed to my benefit and your entertainment.

At CP Kelco, credit goes to Johnny Thylin for supporting the initiation of this project. Many other colleagues were involved in turning it into practical applications: Brian, Anders, Svend, Sarah, Britta, Vicky, Karen, Hanne, Anne, Anita, Jasminka, Anette, Henrik, Flemming, Niels, Søren, Tommy, Vinnie, Lars, Jonathan, Hans Henrik, Morten and all the rest…
And for invaluable support and proof-reading: The one and only Birthe.

At Quality and Technology I'm much indebted to my brothers in arms and colleagues for all the on-topic and especially off-topic discussions and social activities, especially Åsmund, Peter, Thomas, Hanne, Tina, Letícia, Marc, Jakob, Vibeke, Elisabeth, Erik, Birthe, Gilda and Lars & lars.

Thanks finally go to my dear friends: Claus, Peter and all the Jesper's and my parents. Without your support it would not be worth it. And Kim: Thank you for putting the idea into my head – I just had to do it my way at my university!

Christian Bomholt Zachariassen
Copenhagen, June 2007

Table of contents

Summary

The purpose of this PhD study has been to estimate critical process parameters and quality attributes of pectin production, both in-line, on-line and at-line on process liquids and precipitated pectin press cake, and off-line on the final pectin product. The approach taken is that of *Process Analytical Chemistry* (PAC) whose concepts have later been adopted into the broader term *Process Analytical Technology* (PAT). PAT deals with the integration of process analyzers (e.g. spectroscopic instruments) and multivariate data analysis e.g. chemometrics for process control and continuous improvement.

The thesis consists of six chapters followed by one poster and three papers. The first chapter of the thesis gives an introduction and framework for the thesis. The second chapter deals with the basics of spectroscopy followed by an introduction to chemometrics exemplified by data from the three papers. The fourth chapter describes pectin and the benefit from implementing PAT into the pectin production. The final two chapters conclude the thesis and provide a perspective for future developments of PAT and chemometrics.

Paper I describes the development of on-line near infrared based measurement points for the determination of ammonia (and isopropyl alcohol) with focus on on-line control and dosing of the ammonia for the pectin amidation process. **Poster I** and **Papers II-III** describe a novel method developed to measure the carboxylic acid distribution corresponding to the intramolecular ester distribution (i.e. the *blockiness*) of pectin. The presented setup is an off-line method capable of analysing finished pectin. The poster was presented at the 9th Conference on Chemometrics in Analytical Chemistry, Lisbon, 2004 and the papers were all published in the international peer reviewed scientific journal *Chemometrics and Intelligent Laboratory Systems* during 2005-2006.

The research presented in this thesis has revealed a deeper insight and understanding of commercial pectin processing and the structure of pectin. New value has been created through improved product functionality, quality, quantity, and processing capability.

Resumé

Formålet med dette PhD-studie har været at estimere kritiske procesparametre og kvalitetsegenskaber i pektinproduktionen, både *in-line*, *on-line* og *at-line* på procesvæsker og udfældet pektinpressekage samt *off-line* på færdigvare af pektin. Den benyttede indfaldsvinkel er Proces Analytisk Kemi hvis koncepter er blevet adopteret i det noget bredere udtryk Proces Analytisk Teknologi (PAT). PAT omhandler integration af analyseinstrumenter i processen (f.eks. spektroskopiske instrumenter) og multivariat dataanalyse (f.eks. kemometri) til at kontrollere og løbende forbedre processen.

Afhandlingen består af seks kapitler efterfulgt af en poster og tre artikler. Det første kapitel i afhandlingen giver en introduktion og ramme for afhandlingen. Det andet kapitel omhandler basal spektroskopi efterfulgt af en introduktion til kemometri eksemplificeret med data fra de tre artikler. Det fjerde kapitel beskriver pektin og fordelene ved at implementere PAT i pektinproduktionen. De to sidste kapitler konkluderer afhandlingen og giver et perspektiv for fremtidig udvikling indenfor PAT og kemometri.

Artikel I beskriver udviklingen af et *on-line* nær infrarøds-baseret målepunkt til bestemmelse af ammoniak (og isopropylalkohol) med fokus på *on-line* kontrol og dosering af ammoniak i amideringsprocessen af pektin. **Poster I** og **Artikel II-III** beskriver en ny metode udviklet til at bestemme carboxylsyrefordelingen, svarende til den intramolekylære esterfordeling (dvs. om esterne sidder blokvis) i pektinen. Den præsenterede opstilling er en *off-line* metode, der kan analysere færdigvare af pektin. Posteren blev præsenteret på den 9. konference om kemometri i analytisk kemi, Lissabon, 2004 og artiklerne blev alle publiceret i det internationale videnskabelige tidsskrift *Chemometrics and Intelligent Laboratory Systems* i perioden 2005-2006.

Forskningen præsenteret i denne afhandling har åbenbaret en dybere indsigt i og forståelse af kommerciel pektinproduktion og pektins struktur. Ny værdi er blevet skabt igennem forbedret produktfunktionalitet, kvalitet, kvantitet og produktionsevne.

List of abbreviations

A	Absorbance
A	PARAFAC loading matrix
ALS	Alternating least squares (regression)
B	Regression coefficients matrix *or* PARAFAC loading matrix
C	Matrix of concentration profiles *or* PARAFAC loading matrix
CAD	Carboxylic acid distribution
CPP	Critical process parameter
CQA	Critical quality attribute
D	Data matrix (of independent variables) *or* diagonal matrix in PARAFAC2
DA	Degree of amidation
DAc	Degree of acetylation
DAD	Diode array detector
DB	Degree of blockiness
DE	Degree of esterification
DM	Degree of methyl esterification
DP	Degree of polymerisation
E	Residual matrix (of independent variables)
F	Number of components extracted or efficient rank
FIA	Flow injection analysis
FT	Fourier transform
GalA	Galacturonic acid
H	Inner relation residual matrix *or* PARAFAC scaling matrix
HGA	Homogalacturonan
HM	High methylated (pectin)
I	Radiation intensity *or* number of time points
IPA	Isopropyl alcohol (propan-2-ol, 2-propanol)

iPLS	interval PLS
IR	Infrared (spectroscopy)
J	Number of variables or loadings
K	Number of observations (or samples)
LM	Low methylated (pectin)
LMA	Low methylated amidated (pectin)
MCR	Multivariate curve resolution
NH_3	Ammonia
NIR	Near infrared (spectroscopy)
N-PLS	Multi-way partial least squares (regression)
p	Loading vector (of independent variables)
P	Loading matrix *or* PARAFAC ortho-normal profiles
PAC	Process analytical chemistry
PARAFAC	Parallel factor analysis
PAT	Process analytical technology
PC	Principal components
PCA	Principal component analysis
PLS	Partial least squares (regression)
Q	Loading matrix (of dependent variables)
QC	Quality control
RGI, RGII	Rhamnogalacturonan I and II
Rha	Rhamnose
RMSE	Root mean square error
RMSEC/CV/P	RMSE of calibration, cross validation, prediction
S	Matrix of spectral profiles
t	Score vector (of independent variables)
T	Score matrix (of independent variables)
U	Score matrix (of dependent variables)
UV-VIS	Ultraviolet-Visible (spectroscopy)
W	Loading weights matrix (of independent variables)
X, $\underline{\mathbf{X}}$	A data matrix (of independent variables)
y**, y, Y**	Number, vector or matrix of dependent variable(s)

List of publications

Poster I
Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Rasmus Bro, Søren Balling Engelsen, Novel dilution pectin profiling by DAD UV-VIS and three-way analysis, presented at the 9th Conference on Chemometrics in Analytical Chemistry, Lisbon, 2004

Paper I
Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Søren Balling Engelsen, Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amidated pectin production, *Chemometrics and Intelligent Laboratory Systems*, **76** (2005), 149–161

Paper II
Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Rasmus Bro, Anna de Juan, Romà Tauler, Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system, *Chemometrics and Intelligent Laboratory Systems*, **83** (2006), 13–25

Paper III
Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Rasmus Bro, Anna de Juan, Romà Tauler, Multi-way analysis for investigation of industrial pectin using an analytical liquid dilution system, *Chemometrics and Intelligent Laboratory Systems*, **84** (2006), 9–20

Additional publication
Christian B. Zachariassen, An analysis of the innovation climate at CP Kelco, confidential business report, not published

1. Introduction

The purpose of this Industrial PhD study has been to estimate critical process parameters and quality attributes of pectin production, both in-line, on-line and at-line on process liquids and precipitated pectin press cake, and off-line on the final pectin product. The parameters in focus are the concentration of the critical process parameters Ammonia (NH_3) and Isopropyl Alcohol (IPA; which is a common name for propan-2-ol); and critical quality attributes of pectin: Degree of Esterification (DE), Degree of Amidation (DA) and Carboxylic Acid Distribution (CAD). Monitoring these parameters has yielded a deeper understanding of pectin processing conditions. Feedback control of these parameters has created new value through improved product quality, quantity and process capability. Specifically, raw material grading and selection, extraction, amidation and process optimisation have been addressed. Prediction during time of processing enables rapid process feedback and an increase in the proportion of production that meets quality specifications. Included in the dissertation are one poster and three papers. The poster was presented at the 9th Conference on Chemometrics in Analytical Chemistry, Lisbon, 2004 and the papers have been published in international peer reviewed journals. The thesis will give a basic introduction to Process Analytical Technology (PAT), spectroscopy, chemometrics and pectin production with examples of measuring and controlling the pectin production using these tools.

PAT is a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e. during processing by spectroscopy or other process sensors) of critical process parameters and quality attributes of raw and/or in-process materials and processes with the goal of ensuring final product functionality. Near Infrared (NIR) spectroscopy is well suited for rapid at-line and on-line determination of critical process and quality attributes in pectin production. Chemometrics, i.e. multivariate data analysis of chemical systems provides the necessary link between the spectroscopy and the information sought.

The concentration of ammonia to be used in the amidation step of Low Methylated Amidated (LMA) pectin production is a Critical Process Parameter

(CPP) for the control of DA. Likewise the IPA concentration – the IPA/water ratio – is critical for optimal precipitation; therefore both IPA and ammonia are considered CPP's. Furthermore, vast amounts of IPA are used throughout the factory in many processes, and control of IPA concentration and its regeneration in distillation columns play a vital role for the overall energy and cost efficiency of the process. Hence, monitoring and control of IPA process streams have an impact on economy and environment. **Paper I** describes the development of on-line NIR based measurement points for the determination of ammonia (and IPA) with focus on on-line control of the ammonia for the amidation process. Use of chemometric tools such as spectral pre-processing and variable selection is crucial for the success of the application.

Also developed is an in-line and at-line system for the determination of DE and DA in wet High Methylated (HM) and LMA pectin press cake by remote diffuse reflectance NIR spectroscopy. This system measures critical pectin quality attributes directly. The measurement is on the pectin itself, rather than the amidation liquid which can be regarded as a process chemical. The DE and DA of pectin can be controlled by manipulating the peel extraction conditions (i.e. pH, time and temperature) and further on the temperature and ammonia concentration for the amidation reaction. The direct measurement of pectin quality attributes in-line and at-line makes rapid feedback to the process possible. These results have not been published in scientific literature.

Poster I and **Papers II - III** describe a novel method developed to measure CAD corresponding to the intramolecular DE distribution (i.e. the *blockiness*) of pectin. The presented setup is an off-line method capable of analysing finished pectin. Contrary to other published methods, it involves little sample preparation and manipulation, and can easily be upgraded to at-line if desired. The measurements can be used to assess the citrus peel used as raw material (by using samples generated from a test extraction, which is done routinely to measure other raw material attributes on the peel) and as a final product. As presented in **Paper III**, the CAD analysis can be used for HM pectins (with a DE > 50%) only, but can in principle be extended to include Low Methylated (LM) and LMA pectins. Advanced multi-way chemometrics and Multivariate Curve Resolution (MCR) are applied to the data in order to develop predictive models.

Spectroscopy

Improved instrumentation for performing in-line, on-line and at-line spectroscopy has become available during the last decade, and combining input and knowledge gained from several measurement points is now commercially feasible. The instruments yield chemical, physical and structural information about the pectin being produced. Recent advances in NIR detectors, combined with radically improved fibre optics and probe technology make long distance solid-state diffuse reflectance spectroscopy possible using multiplexing NIR spectrophotometers. Thus, it is now possible with one NIR instrument to acquire spectra from pectin at several measurement points on the production line. This can involve designing and inserting probes in harsh process environments. Furthermore, sampling procedures have to be developed for retrieving representative process samples and presenting them in a suitable form to mentioned equipment in order to measure and calibrate the instruments towards pectin structure and functionality measured for instance in the Quality Control (QC) laboratories. The implementation of in-line, on-line and at-line sensors directly on the production line has revealed profound new insights in raw materials for pectin, pectin structure, pectin processing and relationship to pectin functionality in CP Kelco customer applications.

Chemometrics

The development of non-destructive food analysis with the aid of spectroscopic screening measurements goes hand in hand with explorative data analysis/chemometrics. Work has been done to apply, develop and substantiate chemometric algorithms, based on multiple sensor input and key process variables. The available spectroscopic, process and raw material information has been integrated for overall optimisation of production lines in terms of pectin quality and/or quantity. While (single) on-line sensor optimisation on single unit operations is well described in literature, not much information has been published on optimisation of an entire production line.

The synergistic effect of combining advanced chemometric algorithms with advanced spectroscopic sensors has had a great impact on the Danish food industry, which over the last ten years has implemented the methods for at- or on-line quality control to a large extent. This effort has brought the Danish food industry to the technological forefront.

Multi-way models are developed to handle data arrays with more than two dimensions. Data of this sort comes up in e.g. fluorescence spectroscopy or hyphenated methods such as Gas Chromatography-Mass Spectrometry or, as in the case presented in this thesis, UV-VIS spectra recorded during a dilution gradient. Higher order data structures are found in many process applications as well. E.g. the analysis of batch productions runs – organized in batches times local batch time times process variables – will produce a data cube. Multi-way factor models can analyze these data arrays using the smallest amount of parameters, giving robust models and leaving the natural structure of the data intact.

Process Analytical Chemistry and Technology

The purpose of Process Analytical Chemistry (PAC) is "to supply quantitative and qualitative information about a chemical process" and furthermore to utilize such information "…not only to monitor and control processes, but also to optimize its efficient use of energy, time, and raw materials." (Callis, 1987). Those thoughts were formulated 20 years ago as a result of 1) the advances in materials science in the 1970's and 1980's; 2) the improved understanding of process monitoring and control; 3) the emergence of chemometrics as a discipline 4) the availability of chemical and electronic sensors and hardware, especially "microcomputers", which in the terminology of the 1980's are stand-alone computers rather than terminals connected to a mainframe computer.

PAC really matured in the 1990's and the process analytical approach taken was quickly adopted by many industries (Blaser, 1995; Hassell 1998). This led to the introduction of Process Analytical Technology (PAT) where the process analytical part is in focus, and *chemistry* has been replaced by the broader term *technology*. It is no longer significant which kind of sensor provides the signal or which particular process is involved. In some PAC applications the chemistry part did not really apply so PAC became too restrictive. For this reason the use of the abbreviation PAT has prevailed, even though many scientists and engineers are actually doing PAC.

In September 2004, the United States Food and Drug Administration (FDA) launched a new paradigm for the Quality Control (QC) in the pharmaceutical industry. FDA published the PAT guidance document (FDA, 2004), where PAT is defined as: *A system for designing, analyzing, 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. It provides a framework for understanding and controlling pharmaceutical manufacturing processes, an area which is heavily regularized by the authorities, but the standards and practices can readily be adopted by other industries. In fact many industries including the chemical and food industry have been conducting PAT-like activities before FDA formalized it. It marks a paradigm shift from releasing the *final product* based on rigorous post-process quality control by off-line analysis, to releasing products by meeting the specifications in *process design* or by in-, on- or at-line analysis of process data and measurements. The paradigm shift is that "*quality cannot be tested into products; it should be built-in or should be by design."* (FDA, 2004).
In order to achieve this, FDA has recognized four categories of tools to ensure effective and efficient means for acquiring information to facilitate process understanding and provide continuous improvement. The categories are:

- Multivariate tools for design, data acquisition and analysis
- Process analyzers
- Process control tools
- Continuous improvement and knowledge management tools

And FDA continues with the statement: "*…an appropriate combination of some, or all, of these tools may be applicable to a single-unit operation, or to an entire manufacturing process and its quality assurance.*" (FDA, 2004). It is obvious to realize the importance of chemometrics as a multivariate tool for data analysis, the relevance and potential of spectroscopy as (rapid) process analyzers and the need to continuously extract the relevant information in order to provide control and give feedback, not only to the process itself, but also to improve product recipes and facilitate further knowledge acquisition. The relevant information can for instance be the quantification of known or unknown compounds, to ascertain reaction pathways and kinetics, to monitor the integrity of process hardware/equipment, or simply to do on-line feasibility studies.
It has been suggested that PAC does not imply or require the knowledge or management level required by PAT, and that PAC is only used for isolated applications within a process environment. In my opinion, the whole concept was grasped by Callis et al. in their 1987 paper (Callis, 1987).

The modern food industry has many similarities with the pharmaceutical industry. Food products are produced in larger quantities, are moved over longer distances and in longer chains of distribution. This calls for a demand for increasing shelf life but also more rigorous regulations to ensure food safety. Also food products become more sophisticated in their design of texture and consistency, not only to improve the quality, but also to replace the building blocks of food: Carbohydrate, protein and fat with water or air in order to lower the prices or to reduce the energy content of the food, so even more food may be ingested. Public awareness and concern for food safety therefore makes the food industry as tightly regulated as the pharmaceutical industry.

There are notable differences however. It is usually possible and customary to reprocess off-specification food materials. This is rarely accepted in the pharmaceutical industry. This lowers the penalty for producing products off-specification in the food industry. Rigorous testing of pharmaceutical products (raw-materials, in-process products and final products) are therefore carried out in order to minimise waste production. The demand for rapid testing methods is essential to lower production time and storage requirements of products not yet released for further processing.

The system described in **Paper I** is a PAT system by FDA's definition because it provides timely measurements of the ammonia content in the production of amidated pectin during processing. The ammonia concentration is a critical process parameter, as it affects the final product quality of pectin. The estimated ammonia concentration is used directly to dose ammonia in the reactor tanks thus providing a direct feedback into the process. The application takes advantage of all four categories of tools mentioned before: Multivariate tools were used for the overall design and are used in the data acquisition and analysis of the process analyzer chosen. The results are used to control the process, and long term experience has been used to continuously improve the amidation process and increase the knowledge of it.

In Figure 1, from **Paper I**, the essence of PAT is captured. Before the process analyzer was taken in use, the pectin amidation process was controlled by hourly samples titrated by process operators. Based on the desired set point and the result of the titration, the process operator would dose ammonia into the process tank by manually opening a valve, dosing more or less, for a shorter or longer period of time, only based on experience. As the process analyzer (a properly calibrated NIR instrument) became available, measurements were available every 40th second. This made it possible not only to dose ammonia much more controlled and carefully, which was the original

challenge and task at hand. It also revealed that the effect of dosing only once every hour resulted in sharp ammonia peaks 15% above the desired set point , an effect much greater than anticipated.

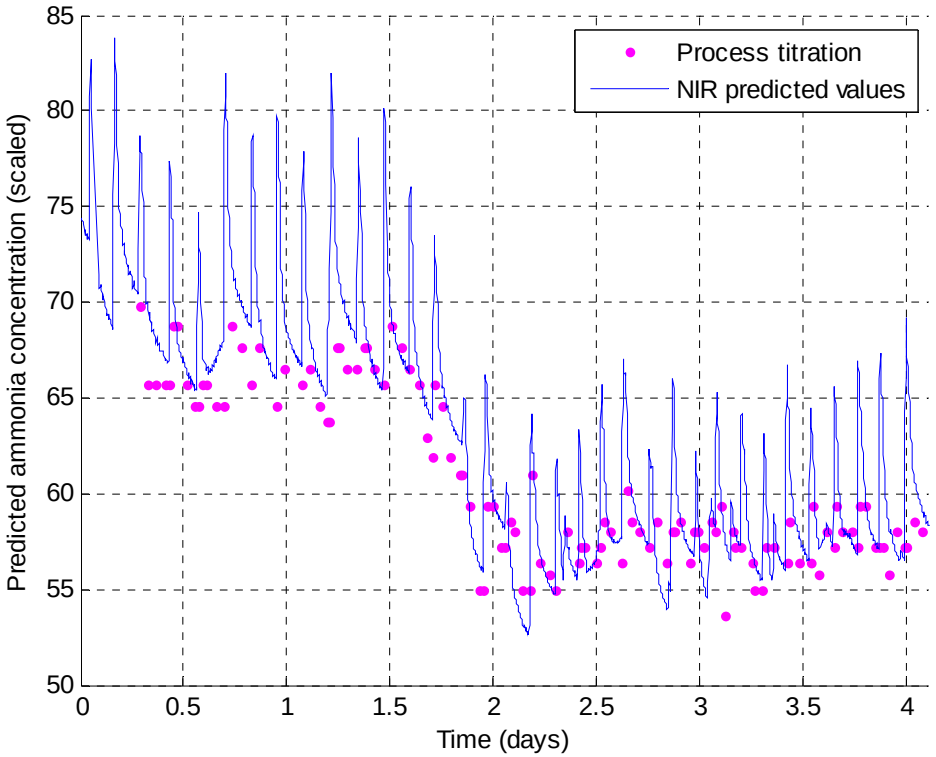


Figure 1 NIR predicted values superimposed on manual ammonia titrations before and after a process set point change taking place at day 2. The hourly titrations do not capture the full dynamics of the manual ammonia dosing process.

The application presented in **Papers II and III** is *not* a PAT system as it is not at-line or on-line in the process in its current setup. It does not provide timely measurements during processing. However, it was deliberately designed as a flow injection system, making it possible to move at-line if desired. It would be a mere question of providing an appropriately conditioned pectin sample to the injection loop.

2. Spectroscopy

Spectroscopy is the study of the interaction between matter and electromagnetic radiation. Spectrometry is the measurement of such radiations as a means of obtaining information about the systems and their components (McNaught, 1997). Electromagnetic radiation is a combination of an electric and a magnetic field. Depending on the energy of the electromagnetic radiation, a number of different processes may take place when the radiation interacts with matter, namely atoms and molecules. The radiation may be transmitted, absorbed, scattered or a combination of all. A molecule or an atom can only absorb energy corresponding to the difference between two energy levels within the atom or the molecule. This will lead to a transition in the state of energy. The allowed states or levels are discrete or quantized and specific for the interacting molecules or atoms. The information can be examined qualitatively or quantitatively and will therefore be of interest in the study of matter.

Electromagnetic radiation can be understood in terms of a wave model as the radiation propagates through a media as waves. Electromagnetic radiation can also be understood as a photon model, where the energy of the radiation is seen distributed as photons, small "packages" of energy, which cannot be divided any further. This duality in the understanding of electromagnetic radiation manifest by the physics of spectroscopy being explained by the wave model introducing terms like wavelength, period, frequency, velocity and amplitude, while concepts from the photon model are employed to define absorption and emission.

The electromagnetic spectrum

Electromagnetic radiation is described by the wavelength, λ, or frequency, υ. The wavelength and frequency of electromagnetic radiation are related through the equation

$$c = \lambda\upsilon \qquad (1)$$

where c is the velocity of light in vacuum, $2.9979 \cdot 10^{8}$ m·s^{-1}. The energy, E, of a photon, which is the quantum charge of radiation is

$$E = h\upsilon \qquad (2)$$

where h is Planck's constant $6.6261 \cdot 10^{-34}$ J·s. Typical units of energy for spectroscopy are the electron volts (eV = $1.6022 \cdot 10^{-19}$ J) or kilo joule per mole (kJ·mol^{-1}). The number of particles in a mole is defined by Avogadro's number N_a, which is $6.0221 \cdot 10^{23}$. This corresponds to the number of carbon-12 atoms present in 12 grams. The wavelength is defined as the length of one passage of the propagating wave of radiation, and for the purpose of chemical spectroscopy is measured in nanometres (nm) which is 10^{-9} m or wavenumbers (σ) which is number of waves per unit, commonly reported as waves per cm (cm^{-1}). The frequency is measured in Hertz (Hz), which is number of wave cycles per second. See Box 1 for an example of conversion of spectroscopic units.

Box 1: Conversion of spectroscopic units:

The radiation perceived as red light in the visible part of the electromagnetic spectrum corresponds to a wavelength, λ, of 700 nm.

Its wavenumber σ is: $1/\lambda = 1/700 \cdot 10^{-9}$ m = 14,300 cm^{-1}
Its frequency υ is: $c/\lambda = 2.9979 \cdot 10^{8}$ m·s^{-1}/$700 \cdot 10^{-9}$ m = 428 THz
Its energy E is: $h\upsilon = hc/\lambda = 6.6261 \cdot 10^{-34}$ J·s*$2.9979 \cdot 10^{8}$ m·s^{-1}/$700 \cdot 10^{-9}$ m
$= 2.84 \cdot 10^{-19}$ J = 1.77 eV = 171 kJ·mol^{-1}

The electromagnetic spectrum (Figure 2) is usually classified in regions by wavelength into electrical energy, radio, microwave, infrared, the visible region (light), ultraviolet, X-rays and gamma rays. For the purpose of chemical spectroscopy, a number of different processes may take place when the radiation interacts with a molecule or atom. In Table 1 a summary of these processes (quantum changes), type of spectroscopy used, and in which part of the electromagnetic spectrum they appear, is shown. The choice of spectroscopy is tightly related to the specific application, the information sought, required precision and accuracy, physical and chemical conditions of the sample and possible interferences. As seen in Table 1, radiation in the UV-VIS region interacts with the electrons on the outer shells of atoms and the bonding electrons in the orbitals of molecules only.

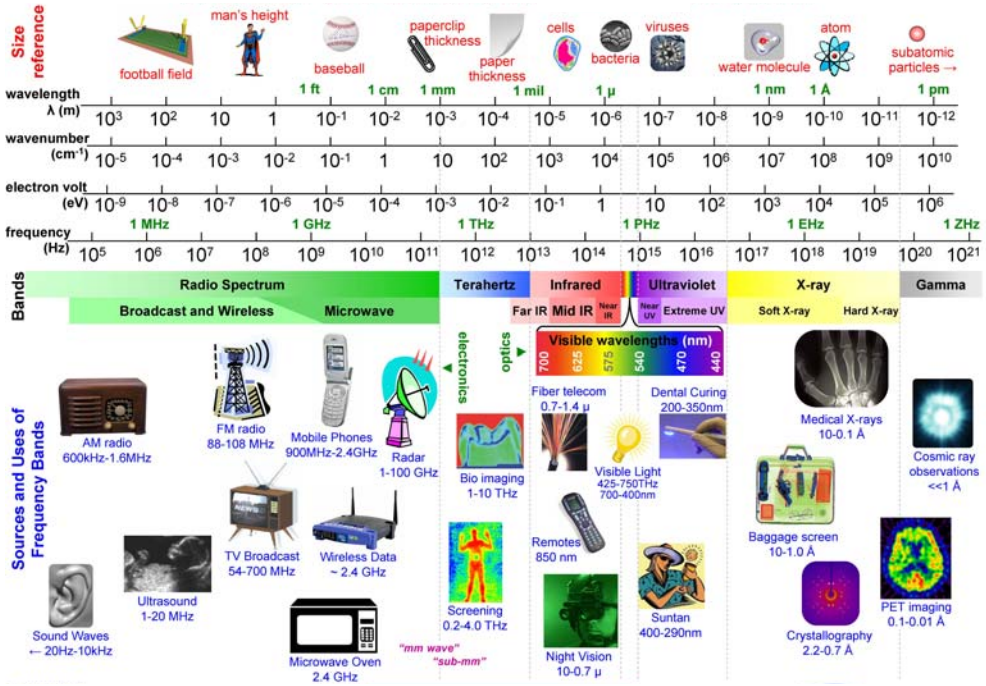


Figure 2 Chart of the electromagnetic spectrum illustrated with radiation sources and applications (reproduced from SURA, 2007).

More energy-rich radiation (Extreme-UV and X-ray radiation) interacts with the inner-shell and non-bonding electrons whereas less energy-rich radiation (near infrared, infrared, microwave, etc.) interacts with the rotational and vibrational states of molecules.

Table 1 List of different types of spectroscopy used and induced quantum change in molecules by electromagnetic radiation of different wavelengths.

Type of spectroscopy	Wavelength range	Induced quantum change
Gamma rays	100 – 0.5 pm	Atomic nuclei excited
X-ray spectroscopy	10 – 0.1 nm	Inner electronic transitions
Ultraviolet (UV)	400 – 10 nm	Outer electronic transitions
Visible (VIS)	780 – 400 nm	Outer electronic transitions
Near infrared (NIR)	2500 – 780 nm	Molecular vibrations
Mid infrared (mid IR)	50 – 2.5 μm	Molecular vibrations
Far infrared (far IR)	1 – 0.05 mm	Molecular rotations
Microwave	1 – 0.1 cm	Molecular rotations
Electron spin resonance (ESR)	100 – 1 cm	Change of electronic spin
Nuclear magnetic resonance (NMR)	10 – 1 m	Change of nuclear spin

Figure 3 shows the quantum changes induced when radiation is absorbed by a molecule in the UV-VIS region, the NIR region, and in the mid IR region. The type of spectroscopy chosen for the application presented in **Paper I** is NIR spectroscopy, while UV-VIS spectroscopy is used in **Papers II and III**.

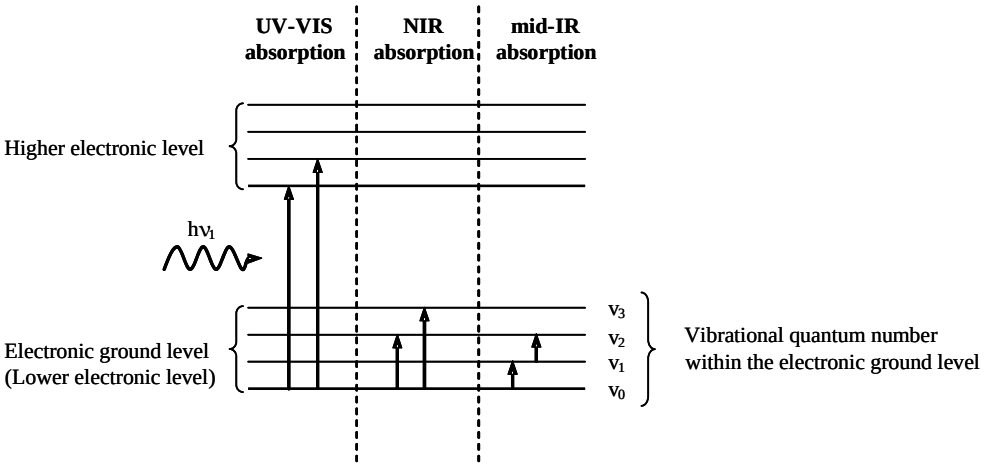


Figure 3 Absorption of light (energy) by a molecule in the UV-VIS region (left), in the NIR region (middle), and in the mid IR region (right).

Basics of quantitative spectroscopy

Quantitative determination of absorbing analytes is approximated by Beer's law, which can be written as

$$A = \log(P_0/P) = abc \qquad (3)$$

where A is the absorbance, P_0 and P the incident and transmitted radiant power, a is the absorptivity proportionality constant, b is the (effective) sample path length and c is the concentration of the analyte (not the speed of light as defined in the previous equations). When a is expressed in the units $L \cdot mol^{-1} \cdot cm^{-1}$, it is called the molar absorptivity and ε is used instead of a. As a replacement for the radiant power P_0 and P the radiation intensity I_0 and I may be used. The power is the energy of radiation in joules hitting 1 m^2 of detector area per second, while the intensity is the joules per second hitting a specific detector.

Beer's law is a first-order approximation that only holds for dilute solutions (as a rule-of-thumb, the analyte concentration should be less than 0.01M = $0.01 mol \cdot L^{-1}$). If several absorbing species are present in a solution, their contribution is cumulative and additive if no interactions between the species

are present. Deviations from Beer's law can also be seen as a result of chemical deviations such as association, dissociation and reaction with solvent. Experimentally, measured absorbance may suffer from instrumental imperfections, such as scattering of light from the surfaces of prisms, lenses, filters, windows, the sample itself or the presence of stray light, all of which contribute negatively to the measured absorbance. There may also be deviations from Beer's law due to the fact that a monochromator will never isolate unique monochromatic light, but a band of wavelengths, which may be more or less symmetric depending on the wavelength range.

Instrumentation

The basic principle of spectroscopy is to convert electromagnetic radiation into an electronic, analytical signal. If the electromagnetic radiation interacts with the matter of interest, then the radiation will undergo transition or change. Spectroscopy is divided in two classes: Absorption and emission spectroscopy. In absorption spectroscopy the radiation – or rather the energy from the radiation – is absorbed by the sample. In emission spectroscopy radiation emitted from the sample is measured.

Instruments for spectroscopy have five basic components (Figure 4):

1) A light source
2) A wavelength selector that isolates a region of the spectrum for measurement
3) One or two sample containers holding the sample and optionally a reference
4) A detector that converts radiation to electric energy
5) A signal processor that amplifies, processes and displays or relays the measurements

Figure 4 Schematic overview of the basic components in spectroscopic instruments. The wavelength selector can be placed before or after the sample container.

In instruments based on the Fourier transform (FT) principle, the wavelength selector is replaced with an interferometer as explained in the following paragraph.

Instrument design

The most inexpensive instrument designs are the single-beam instruments. The sample is measured against a reference (usually the pure solvent without the analyte, if possible) and the sample and reference will have to be switched. In the more expensive double beam instruments the radiation is passed through the sample and reference simultaneously, and is thus to a large degree free from drift in the light source or detector. The wavelength selector is usually placed before the sample to avoid illuminating (and heating) the sample too much. In case of UV radiation, the powerful source may damage or alter the sample. In diode array instruments, the wavelength selector is after the sample to measure all wavelengths simultaneously on the diode array chip. A prism based instrument fitted with a diode array detector is used to acquire the data presented in **Papers II and III**.

All NIR and NIR/IR instruments at CP Kelco, also the one used for collecting the data presented in **Paper I**, are based on the Fourier transform (FT) principle and are manufactured by ABB. A diagram of such an instrument is shown in Figure 5, and on a photo in Figure 6. The heart of an FT-NIR/IR spectrometer is the interferometer. In a standard Michelson interferometer, light from the source enters the interferometer and is divided into two equal beams by a beamsplitter. One beam is reflected towards a fixed mirror, which reflects it back towards the beamsplitter. The other beam is transmitted towards a moving mirror, which also reflects it back towards the beamsplitter. The moving mirror introduces a continuously changing optical path difference between the two beams. As the moving mirror is scanned, the two returned beams interfere with different phases. This creates intensity variations due to interference. At a given optical path difference, the interference is constructive for some frequencies and destructive for others. Because the optical path difference is constantly changing, the various frequencies present in the beam are modulated at different rates. The intensity reading at the detector as a function of the mirror displacement gives an interferogram. ABB instruments are however not fitted with a standard interferometer. They are fitted with a so-called double-pendulum interferometer, where two cube corner mirrors (or reflectors) are mounted on a wishbone scan arm. The scan arms are mounted on a pivot and by moving at small angles this induces a varying optical path difference. When one cube corner advances towards the beamsplitter, the other one moves away from the beamsplitter. The optical path difference corresponds to four times the mirror displacement. Due to symmetry, this type of interferometer is particularly robust towards mechanical distortion (ABB, 2001; Meyer, 2006).

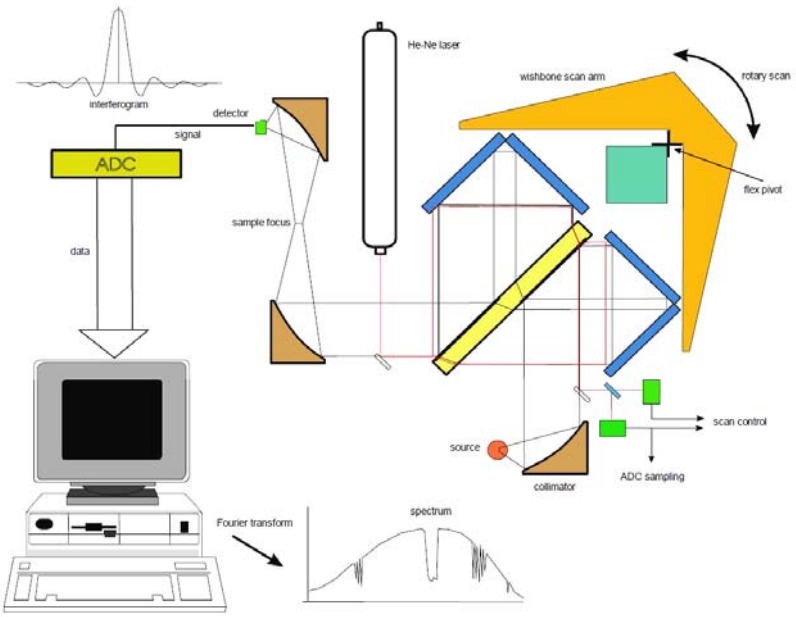


Figure 5 Instrumental setup of an FT-NIR instrument. The beamsplitter is the yellow rod in the centre. An ADC is an Analog to Digital Converter (ABB, 2001).

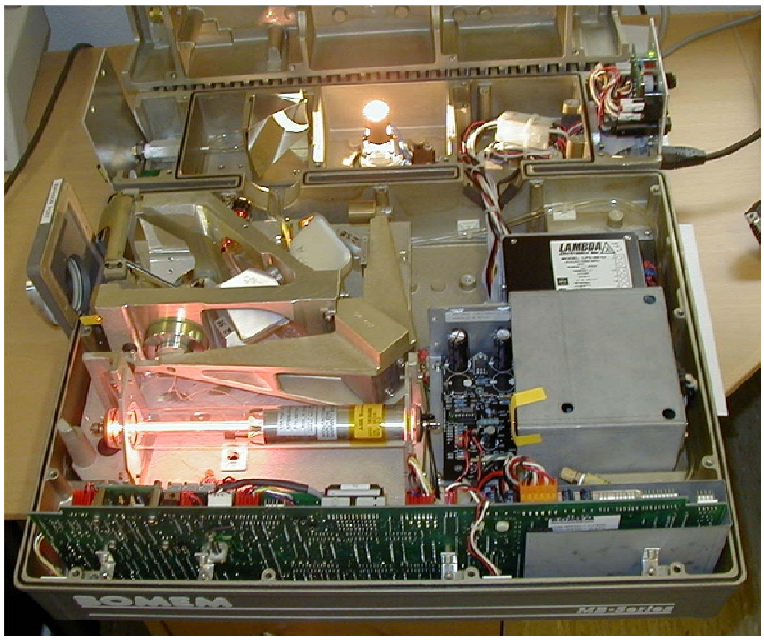


Figure 6 Interior of the ABB MB-160 FT-NIR instrument used in the quality control laboratory at CP Kelco. The light source is at the top centre. Immediately below is the interferometer. Further below is the reference laser ensuring wavelength accuracy. The sample area and detector are not shown.

The spectral information contained in the interferogram is retrieved using a Fourier transformation. A Fourier transformation is a mathematical operation that can be used to relate the interferogram to the spectrum of the sample as it transforms the interferogram from a resolution in time or mirror displacement to a spectrum resolved by frequency.

Sample interfaces

The standard approach is to measure the samples using transmission spectroscopy. The light from the source enters on one side of the sample and is collected on the other side directly across. This is a useful approach for samples which are not opaque, and should be used for a sample path length that does not cause too much absorption; gases and liquids are often sampled this way. In both applications presented in **Papers I to III**, the samples are measured using transmission. In situations where only a single point of entry of the light to the sample is desired or the path length in the sample container is too large, transflection measurement is an option. In transflection, a mirror is inserted in the light path reflecting the light at the same or almost the same angle as the incident beam. The light thus makes a double pass and is collected close to the entry light. In diffuse reflection spectroscopy, the light leaving the sample surface is collected for measurement. It is used when measuring scattering samples or thick or highly absorbing samples e.g. solids, powders or mixtures that do not allow any transmission. The reflected light is collected at an angle to the incident beam usually between 5 and 85 degrees. If the diffuse reflection is not intense enough, the energy can be increased by using diffuse transflection. Similar to regular transflection the light is reflected by a mirror and passes the sample twice, but contrary to regular transflection the light will be diffusely reflected, scattered or transmitted through part of the sample. Diffuse reflection and transflection is of particular importance for NIR spectroscopy.

Sampling modes

The timing of process measurements is crucial. Analyzers can be referred to as being in-line, non-invasive, on-line, at-line or off-line. In-line systems have the sensor integrated within the process line (Figure 7). The sensor may not cover the entire process line or stream, but all process material should have an equal probability of being sampled directly at the point of measurement (Gy, 1998). As measurements are made directly on the process

stream, no delay before analysis is expected and the result can be made available immediately. The analyzer and the sensor can be separated physically; it is the position of the sensor that is important. An example of a sensor placed in-line is when mounted directly in the process, in reactor tanks or in process pipes or tubes.

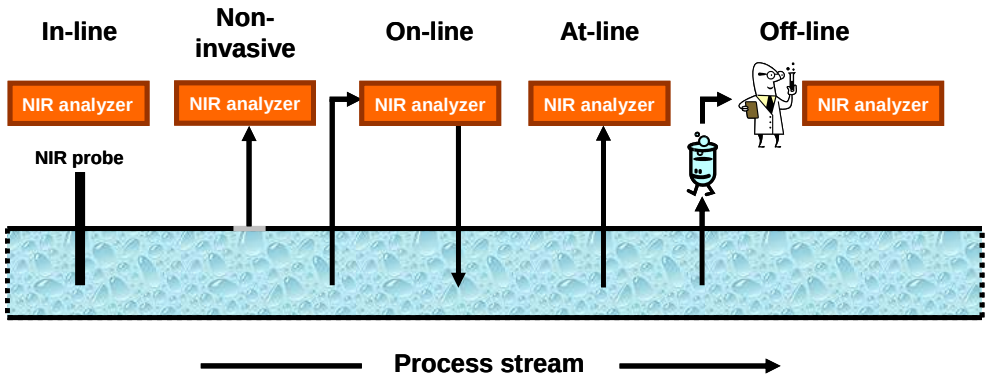


Figure 7 Places on the process stream where spectroscopic techniques (e.g. an NIR instrument) can be implemented. The degree of integration of the NIR analyzer decreases from left to right, whereas the measurement time increases in this direction (modified from Hassell, 1998).

If only a fraction of the process stream is selected prior to measurement e.g. in a side stream sample (fast) loop which has the sensor integrated, the analyzer is said to be on-line. The reasons for choosing on-line rather than in-line are usually practical. It may not be feasible to perform the measurements directly within the process, on a process scale. The physical environment may be unfavourable (subjected to weather conditions outdoors, vibration, sterility requirements, etc.), or the sample needs to be conditioned (e.g. homogenization or the sample needs to be settled for instance to remove bubbles) before measurement. It is also much easier to control the physical environment (e.g. temperature) which would be impractical or costly on a process scale. On-line systems may induce a slight delay e.g. by sample conditioning or by dead volumes. But the sample is always introduced automatically to the analyzer (Blaser, 1995). The results presented in **Paper I** is from an on-line system as amidation liquid is separated from the main process stream in a side stream sample loop and measured before returned to the process tank.

Sensors in in-line or on-line systems may either be invasive or non-invasive. Invasive sensors are sensors which extend into the process and are in direct

contact with the sample, for instance a fixed or retractable optical probe inserted directly in a process line or reactor. Non-invasive sensors do not have (physical) contact with the sample. Examples could be transmittance spectroscopy through viewing glasses, passive acoustic collection of vibrations or remote sensing applications where diffuse light reflected from the sample is collected at a distance. Optical probes embedded rather than inserted into process equipment constitute borderline examples.
At-line analyzers are usually dedicated instruments with the sensor located near or at a process. Sample extraction may be more or less automated, but the sampling and/or sample presentation requires at least some manual handling. The analyzer should not only be near the process in a physical sense, but also in a time-wise fashion. The time taken from extracting the sample to the point when it has been analysed and the results made available or used, should ideally be short and a continuous process, but minor delays can be acceptable. Only a few persons are involved, and most often, the same person extracting the sample is doing the analysis.
Off-line analyzers are situations where the instruments are located away from the production line. Often the time from sample extraction to measurement is not under strict control, timing is not critical and several persons from several departments can be involved, thus passing the responsibility of the sample and its validity. Sample identification, labelling, transport, queuing, prioritizing and registration become important parameters. An example is samples extracted at the process line and measured in a centralized laboratory environment e.g. a Quality Control (QC) laboratory. There can be several good reasons for not committing instruments at-line or on-line. The instrument might be expensive and therefore shared among several applications. The instruments may not be robust or otherwise suitable for a process environment. It may be an advantage to analyse samples in bulk, or specially trained or otherwise qualified personnel may be required to analyze the sample (Vidrine, 2000).

Principles of visual spectroscopy

The UV-VIS region is not defined unambiguously. Most sources will define the UV region as starting between 1 nm and 200 nm and ending between 380 nm to 400 nm. The visible region is defined as starting between 380 nm and 400 nm and ending 750 nm to 780 nm. For the purpose of food spectroscopy, the UV region will be defined in this text as ranging from 180-400 nm, and the VIS region as ranging from 400-780 nm. This definition of the UV region

excludes what commonly is referred to as the Extreme or Far-UV region, which could be defined as ranging from 10-180 nm. A limited number of functional groups exhibit absorption in the UV-VIS wavelength range. Such active functional groups are called chromophores. Examples of chromophores are aromats, conjugated alkenes, amines, carboxylic acids, esters, amides, thiols, polyenes, peroxides, nitrites, ketones and aldehydes.

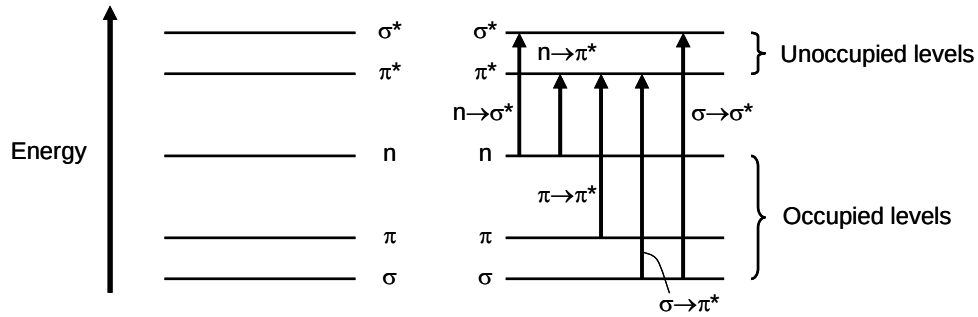


Figure 8 Energy levels of molecular orbitals (modified from Pavia, 2001).

When a sufficiently energy rich photon is absorbed, electrons in appropriate molecular orbitals can be excited from the occupied bonding levels to the unoccupied non-bonding levels. Figure 8 illustrates this process schematically; while Figure 9 lists some commonly observed orbital changes in various compounds.

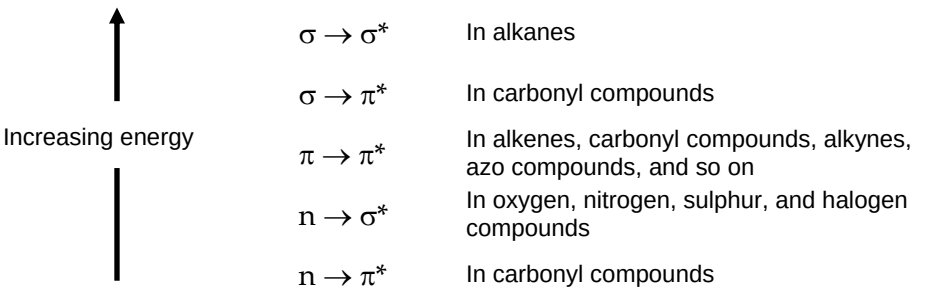


Figure 9 Observed molecular orbital changes in different compounds (modified from Pavia, 2001).

Principles of infrared and near infrared spectroscopy

Molecules are in a constant state of vibration, and each bond is having its characteristic stretching and bending frequency. The characteristic stretching and bending mode vibrations for the $–CH_2–$ group are shown in Figure 10. Molecular vibrations are not purely harmonic. As a result infrared absorption can also occur at multiples of the fundamental frequency. Particularly multiples of the C-H stretch frequency found in all organic molecules (at multiples of ~2900 cm^{-1}), are often used for infrared analysis. These frequencies, which occur above 3000 cm^{-1}, are in the near infrared region of the spectrum. Near infrared spectra may also exhibit absorption due to combinations of overtones and of other fundamental vibrations. Usually, typical regions of the spectrum are selected to be used for analysis.

The choice of spectral regions is normally determined by the limitation of sampling. In the mid IR region, including the C-H stretch region (4000 cm^{-1} to 400 cm^{-1}), the pathlength through samples consisting of solids or semi-solids are generally in the range of a few micrometers to 0.1 mm. In the mid IR region, absorption bands due to different functional groups are least overlapping. Here it is possible to detect small concentrations in mixtures of materials by increasing the pathlength such that the bands of the main components are highly absorbed while the weak features of the low concentration component occur between the dominating main components. Some selectivity in the wavelength region is typically required of some of the functional groups of the low concentration component.

In the region where combination frequencies occur (5000 cm^{-1} to 4000 cm^{-1}), the pathlength is usually less than 1 mm. In the 1st overtone region (7000 cm^{-1} to 5000 cm^{-1}) the pathlength is typically 3 to 5 mm. In the 2nd overtone region (10000 cm^{-1} to 6000 cm^{-1}, still in the NIR region) the pathlength is often as much as 10 mm. The longer possible pathlength in NIR than IR makes it possible to have a high degree of flexibility in the sampling so little or no sample preparation or conditioning is required. Fibre optical cables are virtually transparent in the NIR region and can be used over long distances, making it possible to separate the instrument from the sampling area. As IR, NIR measurements are rapid, non-destructive and reproducible. Therefore NIR is well suited for PAT applications. In Figure 11 a chart of common NIR functional groups in wavenumbers can be seen. NIR can fingerprint all organic molecules and are ideal for pharmaceutical or food analysis.

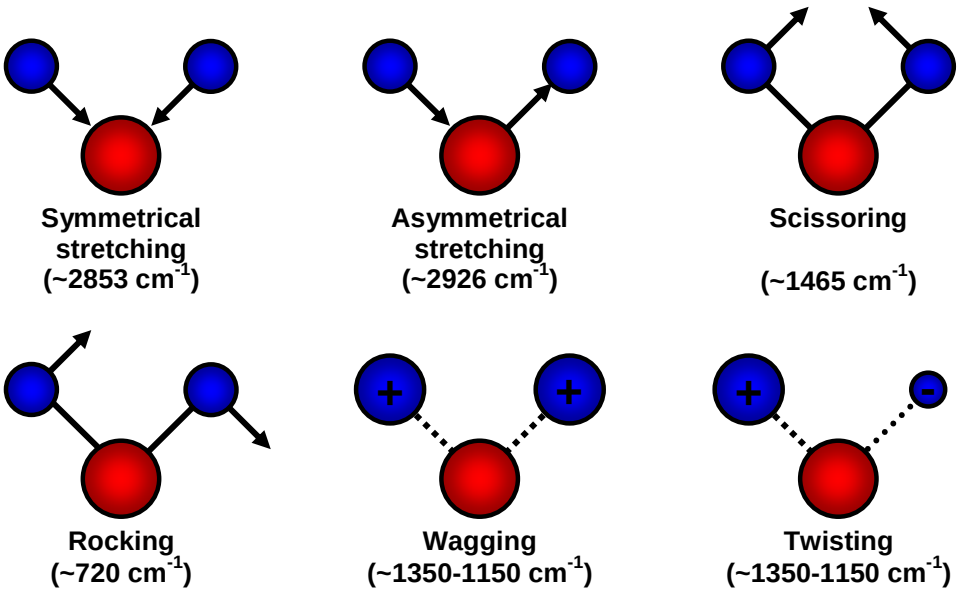


Figure 10 Modes of vibration of a $-CH_2-$ group and their approximate frequencies. The non stretching vibrations are also referred to as bending and deformation vibrations (modified from Silverstein, 1991).

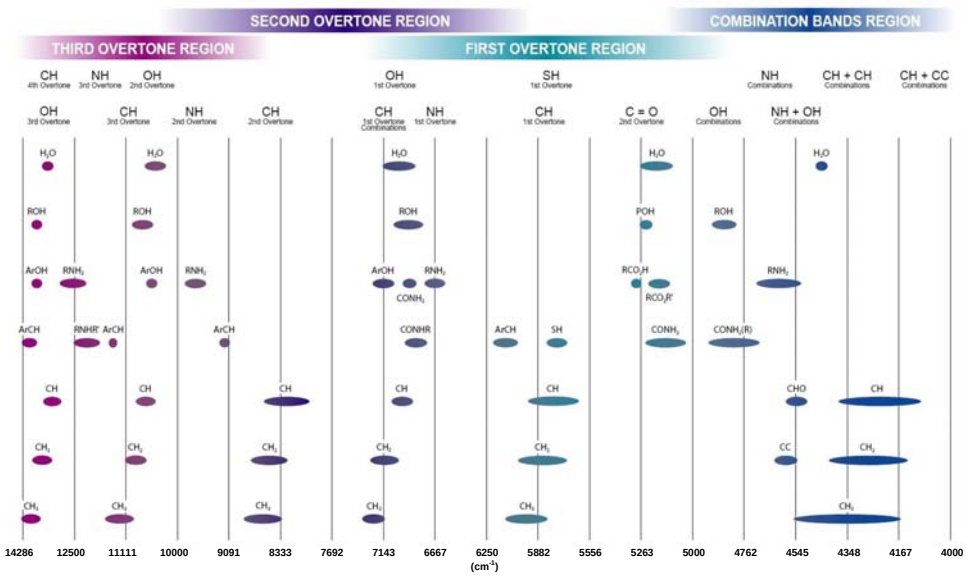


Figure 11 Chart of absorption bands in the near-infrared region (reproduced from ASDI, 2007).

3. Chemometrics

The world around humans is a continuous evolution of complicated phenomena. The nature that surrounds us is multivariate in the sense that all phenomena depends on multiple factors. Some factors are immediately recognizable to humans, as the brain is capable of processing data multivariately. Other factors are hidden – latent – because their causalities are complex and not immediately presentable nor interpretable. The use of language, which is a rather new innovation in light of the human evolution, is an expression for (a simplification of?) the thoughts and feelings, that interrelate humans. These relations can be attempted to be reproduced by more or less refined expressions as gesticulations, speech, writing, poetry and art with varying degrees of success. The written language that humans have developed, where the spoken word is recorded with simple symbols in the shape of letters and signs are even further away from the inductive thought process. There are obvious limitations when variables and objects are reduced to data tables. It is difficult to depict objects in three dimensions, let alone four. To portray eight variables (dimensions) in a table is conceptually difficult. Traditionally this is solved mathematically by computing figures of merit (e.g. mean, deviation, skewness, etc.) for each variable, or graphically by inspection of one variable at a time. This however puts emphasis on the individual variables themselves and could lead to wrong conclusions because the *relationships* between the variables, which may be much more interesting for the problem at hand, is ignored. In chemometrics emphasis is given to the covariation among variables. Chemometrics originates from *chemeia* (Greek, chemistry) and *metros* (Greek, measure).

Rather than using deduction – by disregarding the overall impression which is not fully understood, and breaking nature down into subunits which can be modelled using hard modelling – chemometrics is a tool for induction. Empirical data can be analysed *without* constructing a hypothesis of the relationship between the variables prior to commencing. This makes it possible to perceive entire soft modelled phenomena (latent structures) rather than individual subunits (Munck, 1998; Munck, 2006). The definition of chemometrics is still under debate, but the ICS (International Chemometrics Society)

conference chair offers the following definition: *Chemometrics is the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods.*
Chemometrics to a high degree makes use of projection based data analytical methods, which make it possible to reduce the dimensionality of complex problems, if they can be expressed in data tables or arrays. This enables an intuitive graphical presentation of even large data sets, and makes it possible to understand and interpret the latent structures otherwise hidden in the measurements. Therefore, chemometrics relies heavily on the use of computers and graphics. Microcomputers were becoming available in the seventies and eighties, and since then personal computers have been mass produced and fitted with memory and processors that have increased their capacity and speed thousand fold since the beginning of the 1990's (Levy, 1988). Multivariate data analysis requires a lot of computing power, and even though the principles were described in the start of the twentieth century they were not commonly employed before the seventies. To illustrate the leap in computer power: One of the most popular chemometric software packages, The Unscrambler, required an IBM PC with 386, 40MHz processor, 4Mb RAM and 1,5Mb free hard disk space (for the programme itself) in the 5.5 DOS version available in the mid-nineties. Today this has increased roughly thousand fold. The complexity of chemometric problems that can be reasonably solved have likewise increased.
The purpose of multivariate data analysis can be divided into three main categories:

Exploratory data analysis: Screening of data, projection on latent factors and identification of clusters (groups) and outliers (single observations) in the data set.

Classification: Discrimination between groups in the data set and/or pattern recognition. Pattern recognition methods do not require prior knowledge to groups or classes in the data. Classification and discrimination methods require predefined classes and will be allocating membership of new objects (samples) into these predefined classes. The classes might be very well described using exploratory data analysis.

Regression and prediction: The purpose of regression methods is to link two groups of variables. If the variables can be divided into *independent* and *dependent* variables, a regression or calibration from the independent to the dependent variables can be estimated. Independent variables are the quanti-

ties measured or computed, with values chosen by exploratory data analysis or by design when performing the calibration process. It is usually supposed that the independent variables carry negligible error. The dependent variables are the quantities measured or computed as the function of the independent variable. The dependent variables are usually subject to errors and deviations. The main purpose of the calibration is usually to predict or estimate the dependent variables from new sets of measurements of the independent variables.

Analytical chemistry and chemometrics

An example of the use of traditional analytical chemistry based on deduction, and a hard mathematical model being developed, is the previously mentioned Beer's law (Equation 3). This is a hypothesis of the interaction between the incident and transmitted radiant power, the absorptivity constant of the analyte(s), the sample path length and the concentration of the analyte(s). This equation can be applied to observations made, and the parameters can be estimated in experiments. But as mentioned, the equation only applies to dilute solutions. It does not take into account any interaction between the absorbing species, other electrostatic interactions, and last but not least, interferences. Especially in nature, interference from many sources may be present, known or unknown. The traditional way of eliminating these have been selection, to make the analytical signal as selective as possible. This usually involves a lot of unit operations on each sample such as extraction, precipitation, masking and so on. The result is *one* expensive signal per sample, e.g. weight of precipitate, absorption at 420 nm, specific rotation of polarized light, etc. *Low speed – high cost* is the trademark of most wet-chemistry methods. There are few variables and many samples. The data structure is said to be *vertical* and classical statistics is suited to this structure. The parameters can be estimated; means, standard deviations and confidence intervals can be calculated.

Many measurement methods are being transferred from the traditional wet chemistry methods to instrumental analytical methods. New computer controlled instruments are capable of producing large quantities of data in short time. The instruments are to a higher degree fitted with auto samplers ensuring around the clock capability. The need for sample preparation is minimized and the analysis time per sample is commonly short. Often the capital cost of the instruments is high but the cost in use for each sample is typically low. So, after the initial investments, instrumental analysis is often

high speed – low cost. Examples of such instrumental analysis are IR, NIR and UV-VIS spectroscopy, Mass Spectrometry (MS), Nuclear Magnetic Resonance (NMR) spectroscopy, Image Analysis, Gas Chromatography (GC) and High Performance Liquid Chromatography (HPLC). In these examples, the number of variables measured is usually larger than the number of samples, and the data structure is said to be *horizontal*. Often, however, the measured variables are not selective with regard to the problem at hand. They are only indirect measurements, but by measuring hundreds, thousands or even millions of indirect variables it may still be possible to establish a correlation to the desired properties. The key is multivariate data analysis. Chemometrics is tailor-made to analyze horizontal data structures e.g. 50 samples and thousands of variables. The catch is that the relative "safety" of classical statistics is lost. The estimations of uncertainty, standard deviations and confidence intervals are less well defined. This is circumvented by using alternative measurements for the reliability of the estimated results, approximated under certain assumptions, which will be elaborated upon in the section about validation. All in all, the combination of modern analytical instruments combined with chemometrics, makes it possible to measure directly on nature with simple or no sample preparation, and relate the measurements to the desired properties.

Order of chemical measurements

Chemical measurements can be of different order. For instance the earliest analytical instruments were zero order instruments. They produce a single data per sample measured; a zero order tensor. Examples are pH electrodes, temperature sensors, absorbance at one wavelength, etc. For instance the amount of total organic carbon in source water and drinking water can be determined by specific UV absorbance at 254 nm (Potter, 2005). The method is capable of measuring a single analyte (carbon) in known mixtures before applying the method to an unknown mixture. To ensure selectivity of the signal, the sample is acidified and the inorganic carbon is removed prior to analysis for organic carbon content. There is no way to detect errors due to the presence of interferents (e.g. inorganic carbon) with a zero order instrument. It is not even possible to tell if there is an interferent present.

First order instruments produce a vector of data for each sample run. Examples are multi-wavelength spectroscopy (MS, NIR, IR, UV-VIS absorbance, etc.), chromatography (e.g. liquid or gas chromatography), capillary electrophoresis, sensor arrays, etc. The application presented in **Paper I** is using a

first order instrument and gives first order measurements. The advantage of first order measurements compared with zero order measurements is that the presence of interferents can be detected. But it can only be corrected for, if the interferents are known at least to the level that they can be spanned in the calibration set, and therefore part of the calibration procedure. Second order instruments produce a matrix of data for each sample run, for instance excitation-emission fluorescence spectroscopy or hyphenated instruments obtained by combining two first order instruments. Examples can be combinations of Gas Chromatography and Mass Spectroscopy, (GC/MS); High Performance Liquid Chromatography (HPLC) with multi-wavelength UV-VIS absorbance (HPLC/UV-VIS), etc. The presence of unknown interferents can be detected and mathematically eliminated using second order chemometric methods. Thus, interferents do not have to be part of the calibration standard. This is referred to as the so-called "second order advantage" (Booksh, 1994; Bro, 1998; Prazen, 1998). Also, some forms of instrumental run-to-run drift can be corrected. Even higher order measurements are possible, for instance fluorescence excitation-emission-time decay measurements or GC/MS/MS. As the information content of measurements increase going from zero to higher order analysis, so does the complexity of the data. This limits or complicates the visualization and choice of data analysis methods available, as will be illustrated later in this chapter.

When many samples are analyzed together the *data* dimensionality will increase. Many samples with zero order measurements will give a data vector to be analyzed. Likewise, samples with first order measurements give a matrix (2-dimensional array) of data, where samples are one of the dimensions. Confusion can arise of the fact that the data dimensionality can be higher than the dimensionality (order) of the measurements.

The setup used in **Papers II and III** features a Continuously Stirred Tank reactor (CSTR) where the analytes (pectin and dye) is retained while chemical reactions between the analytes occur. This setup has features from Flow Injection Analysis (FIA; Růžička, 1988) and HPLC. The concentration of pectin changes during the sample run and a first order instrument, a diode array detector (DAD), records the UV-VIS spectra during the dilution. The sample matrix will have the format retention (or dilution) time × wavelength. When several samples are analyzed as a whole, a three-dimensional array will be the starting point.

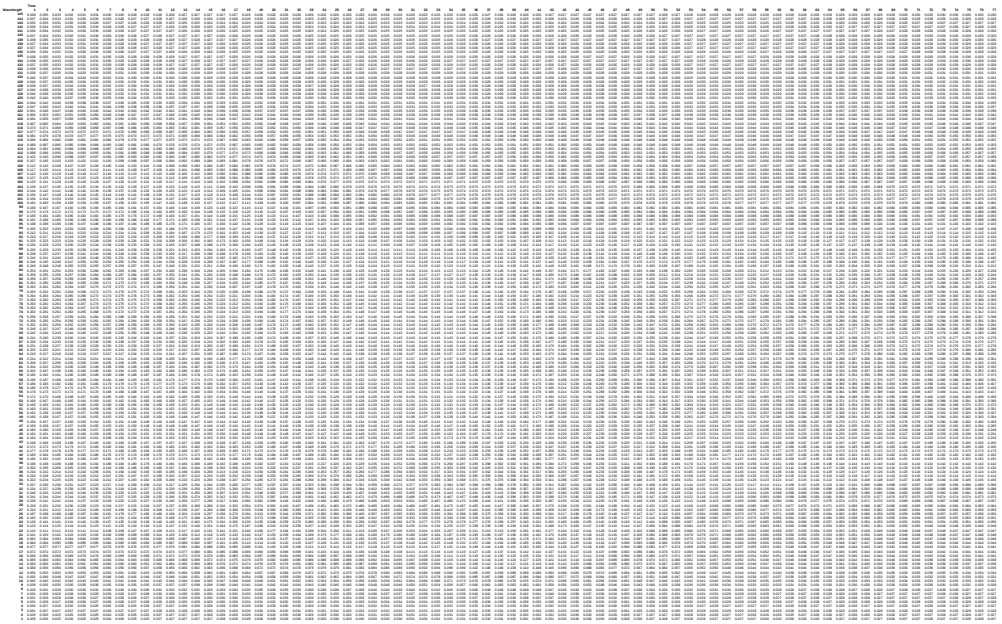
The setup used in **Paper I** does *not* give two-dimensional data although the timely measurements of the on-line system could be regarded as a dimension. It can be argued that *if* the measurements could be split into different

batches *and* a separate batch would constitute a sample as a whole, then a matrix of local batch time (of individual measurements) × wavelengths can be constructed. But it is not the purpose of the application to analyze the ammonia content or consumption of a batch as a whole, it is rather the purpose to make a determination of the ammonia content in each individual measurement, independent of other measurements. Therefore, the individual measurements constitute a sample. Likewise, combining two sensors which cannot meaningfully be arranged in a matrix does not give a higher order measurement. For instance combining a NIR and IR sensor on chemical measurements does not give a second order measurement; rather the two sensors are extensions of each other. A collection of several zero order descriptors does not constitute separate dimensions unless they directly influence each other. The Excitation-Emission Matrix (EEM) of fluorescence measurements *is* of second order, as the emission is a direct property of the excitation wavelength.

Visualization of second order data

The complexity and sheer size of second order data can be illustrated with the data presented in **Papers II and III**. The raw data matrix from one sample consists of 333 spectra, each with the absorbance recorded from 2048 specific wavelengths. This translates to 333 time points × 2048 absorbencies, in total 681,984 data points. This can be reduced by cropping wavelengths and spectra with no relevant information and by reducing the data by averaging wavelengths and time points. The dataset presented in **Paper III** consists of 33 samples measured in triplicates, and each sample matrix is sized 77 time points × 145 wavelengths. This accounts for 99 sample runs of 11,165 data points per sample run – still more than one million data points in total.
Table 2 illustrates the point made in the introduction to this chapter, that letters and numbers may not be the most efficient way of communicating information. The table shows one sample run of a specific sample (please refer to **Papers II and III** for specifics regarding the sample and sample naming). Clearly there is a need for multivariate visualization and data analysis. The origin of the data is UV-VIS spectra recorded from a Diode Array Detector. It records the transmittance expressed in absorbance units. The absorbance can be regarded as a function of wavelength and spectra number which corresponds to the retention or dilution time of the system. This can be plotted as a three-dimensional system with (Time, Wavelength, Absorbance) on the three axis, a so called spectral landscape.

Table 2 The data matrix from sample "B36.0 R0.0" (already reduced to 1.6% of the size of the raw data collected). The numbers are the absorbance recorded as a function of time and wavelength channel. The actual numbers may not be legible due to printer resolution limits or picture compression in the document.



To visualize this on a flat surface is already a challenge, but is partly feasible through the use of perspective, overlap and colour. The bottom left part of Figure 12 shows a pseudo three-dimensional view of the data matrix. The individual spectra as recorded by the DAD (following the before mentioned data reduction) can be seen on the top left part of the figure. The wavelengths (on the x-axis) are recorded simultaneously and presented as a vector with absorbance on the y-axis. This is the traditional way of viewing UV-VIS spectra. The first spectra recorded are coloured blue and the last spectra are coloured red. On the top right side of the figure is a plot of each individual wavelength as a vector with time on the x-axis and absorbance on the y-axis. Wavelength 1 is coloured blue and wavelength 145 is coloured red. On the bottom right side of the figure is what will be the preferred way of presenting the three-dimensional data. It is reduced to a two dimensional plot with time on the x-axis and wavelength on the y-axis. It does *not* have (false) perspective or overlap, and the absorbance is translated to colour. All parts of the spectra are readily seen and can be appreciated by all with a normal colour vision – it will prove a challenge for those who are colour blind; my apologies!

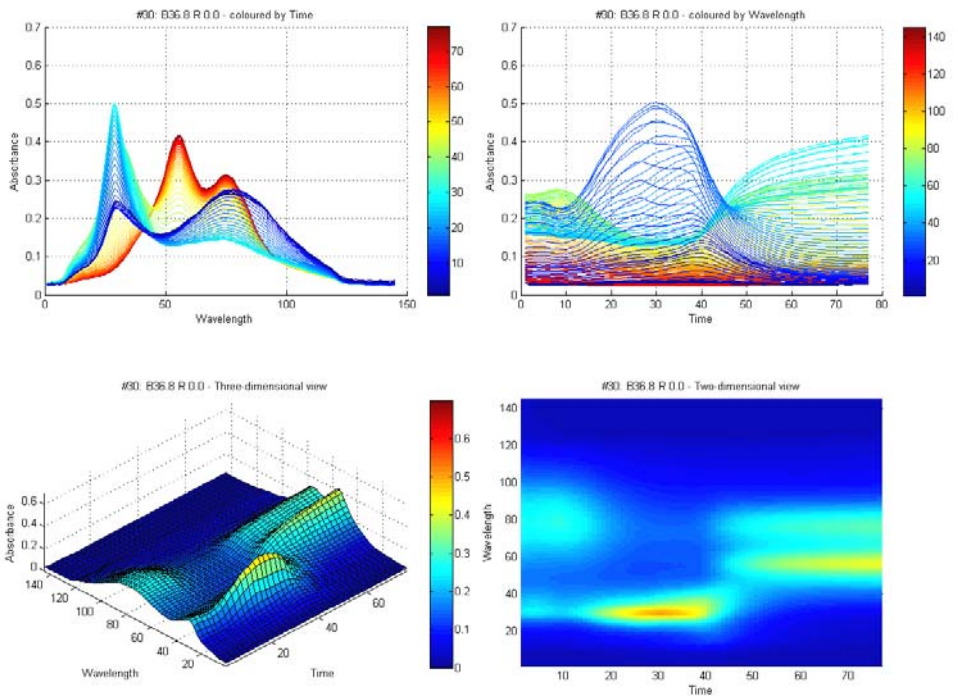


Figure 12 Different ways of illustrating the data from one sample run. The top part of the figure shows to the left, spectra coloured by time; on the right time profiles coloured by individual wavelengths. The colour absorbance legend is shared between the bottom parts of the figure showing pseudo three-dimensional and two dimensional views with absorbance values translated into colour. Absolute wavelengths and times cannot be disclosed for reasons of confidentiality.

In literature it is the convention in the western culture to start reading in the top left corner and proceed across to the top right corner and work this way across to the bottom. In mathematics it is the convention to have origin of a coordinate with the primary axis (the x-axis) going horizontally across from left to right, while the secondary axis (the y-axis) is extending vertically up. The origin is thus placed in the lower left corner, if only the first quadrant of a cartesian coordinate system is considered. It is chosen by the author to adopt the mathematical convention. Furthermore, it is chosen to have "Time" on the primary x-axis as then progression in time (and spectra) can be seen going from left to right in agreement with both literature (in the western culture) and mathematics. The first wavelength is therefore placed on the bottom row and the last wavelength is placed at the top. This is in agreement with the mathematical convention and opposed to the literary

convention. It is for this reason that Table 2 has *wavelength* number 145 in *row* number 1 and declining towards wavelength number 1 in row number 145. The absorbance is seen as a function of Time and Wavelength, so Absorbance is on the third axis (the z-axis) and commonly reduced to representation by colour in two dimensional plots. This is, by the way, a common way to depict heights on physical geographical maps. Colours are still subjectively perceived, shown differently on various monitors and printed differently on individual printers. If not otherwise specifically mentioned, the absorbance colour scale goes from 0 (deep blue) to 0.7 (dark red) and the colour scale in words (if such is to be used) will be an 8 level scale of dark blue, blue, cyan, green, yellow, orange, red and dark red although reference to absorbance in numbers rather than colours will be preferred for reasons of objectivity. The colour scale corresponds to the colours of the visible part of the electromagnetic spectrum except that the violet part is omitted. In this way, the numbers in Table 2 have been translated to a two-dimensional image (or landscape) where absorbance is translated into colour and spectral peaks or heights with high absorbance will have a redder colour.

Unfolding of second order data

The dimensionality of the data can influence or restrict the type of chemometric methods available (Escandar, 2006). For the analysis of several samples with first order measurements, the data can appropriately be arranged in a matrix with samples in the rows and the measurements in the columns. The matrix can then be analyzed with common chemometric methods such as Principal Component Analysis (PCA; Wold, 1987; Jackson, 1991) or Partial Least Squares Regression (PLS; Wold, 2001). Likewise one sample with second order data can be analyzed with the before mentioned chemometric methods or Multivariate Curve Resolution (MCR; Tauler, 1995a) can be applied. Several samples with second order measurements can be organized in a three-way array exemplified with the data from **Papers II and III** on the bottom right part of Figure 13. Three-way arrays can be analyzed using three-way or multi-way chemometric methods such as PARAFAC (Bro, 1997), PARAFAC2 (Kiers, 1999) or multi-way Partial Least Squares regression (N-way PLS or N-PLS; Bro, 1996a). The different chemometric methods will be presented in more detail later in this chapter. It is however also possible to unfold a three-way or multi-way array into a two or one dimensional array by a process known as unfolding or augmentation in general, or more specifically matricizing. This makes use of the more common two-way chemometric methods possible.

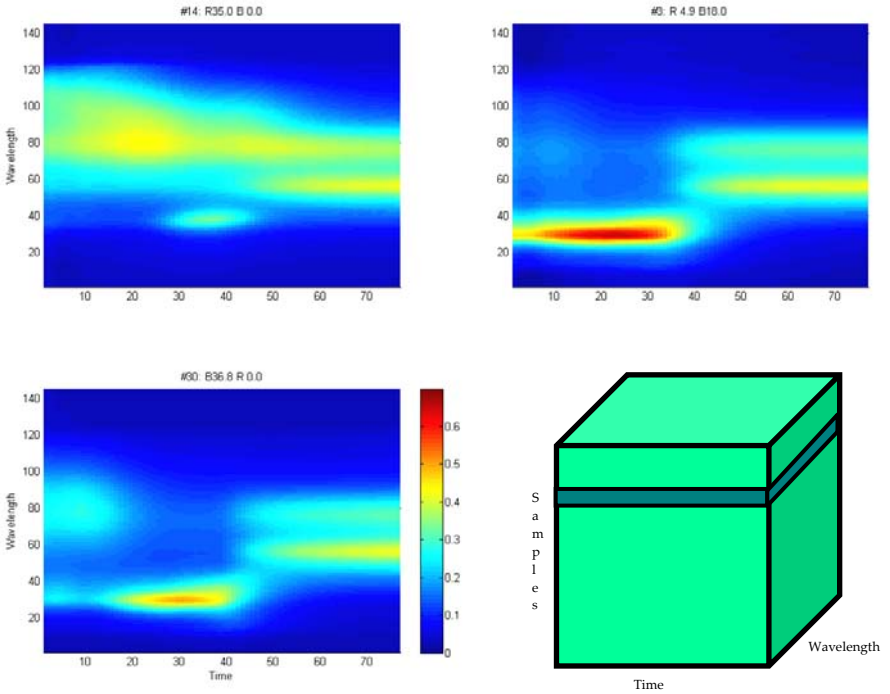


Figure 13 On the bottom right is a three-way data array (a cube) of data. Each slab consist of the two-dimensional matrix of Time × Wavelength. Such three examples of sample slabs are illustrated on the top and bottom left side of the figure.

Unfolding a multi-way array is not trivial. For instance a three-way array can be unfolded in six different ways each looking at a different type of variability. Three of the six ways will be identical result-wise when analyzed by PCA or PLS, as they will only have their rows or columns rearranged (Westerhuis, 1999), but they will appear differently for interpretative purposes. As an example, three samples from **Paper II** have been matricized in two ways (Figure 14) using the convention in the paper by Anna de Juan et al. (de Juan, 2003); therefore in this plot wavelength is on the x-axis and the time of the unfolded landscapes are numbered from the top left corner of the y-axis. A third way of matricizing is vectorizing the individual samples. Samples can then be put in the rows of the matrix and the vectorized measurements in the columns. Again this can be done in two ways, keeping either the individual spectra together and augmenting the spectra recorded at various times, or the other way, augmenting the wavelength channels. The PCA and PLS models will remain the same, but the interpretability and modelling focus will change. If spectra recorded at specific times should be excluded from the modelling – preference should be in unfolding in the way

that keeps the individual spectra together. As an example one sample from **Papers II and III** is vectorized both ways as seen on the bottom right part of Figure 15. The length of the vector is 77 × 145 = 11,165 data points in total.

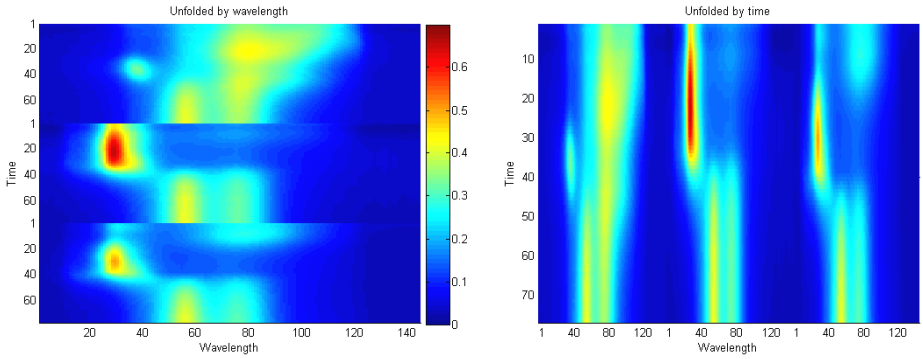


Figure 14 Two ways of unfolding three sample runs from a cube to a matrix with the wavelengths on the x-axis. On the left is the column-wise augmented matrix and on the right is the row-wise augmented matrix.

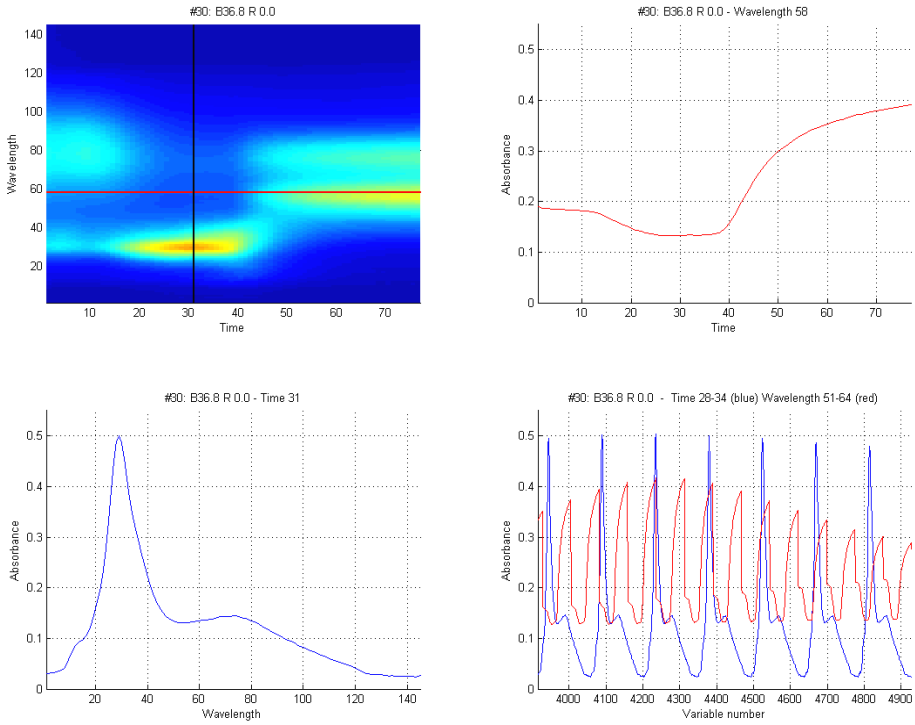


Figure 15 The sample on the top left side is vectorized either by augmenting vertically, keeping the individual spectra together, or horizontally by preserving the wavelength channels. Part of the resulting vector can be seen on the bottom right side of the figure as the blue and red line, respectively.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA), also called Singular Value Decomposition is one of the most commonly used methods in chemometrics (Wold, 1987; Jackson, 1991; Beebe, 1998; Umetrics, 2006). PCA is capable of decomposing observations (made on a number of objects or samples) into a structural part and a noise part. If it is assumed that:

$$\text{Observations} = \text{Data structures} + \text{noise} \quad (4)$$

Then the noise is imposed on data structures and the purpose of a PCA is to split the observations into a structural part and a residual part containing the noise. If the observations can be organized in a matrix $\mathbf{X}$ (of size $K \times J$) with K observations (or samples) in the rows and J variables in the columns (see Box 2 for the notation used), then the mathematical rank or dimensionality of the matrix is defined as the minimum of K and J. But if the variables (or samples) exhibit some degree of covariation, the effective rank of the matrix is said to be lower than the minimum of K and J. The structural part with the systematic variation forms a subspace of $\mathbf{X}$. The chosen subspace has F dimensions corresponding to the effective rank. F is thus smaller than the minimum of K and J.

Box 2: Notation used in the thesis:

Running indices have lower-case italic fonts: x
Scalars have upper-case italic fonts: X
Vectors have lower-case bold fonts: $\mathbf{x}$
Matrices have upper-case bold fonts: $\mathbf{X}$
Multi-way matrices have upper-case, underlined bold fonts: $\underline{\mathbf{X}}$

The transpose is marked with a superscript T: $\mathbf{X}^T$
Elements in matrices can have subscript italic running indices for clarity: $\mathbf{X}_{kj}$, implying that $\mathbf{X}$ is of dimension $K \times J$.

The notation in the spectroscopy chapter deviates from this due to conventions established within spectroscopy.

In order to split the observations in **X** into a structured part and a residual part (commonly referred to as the residuals) containing the noise, principal components (PC's) are extracted from the observations. PC's are constructed from *scores* and *loadings*. Loadings define the relationship between the original variables measured and the PC's. For every PC there are J coefficients corresponding to the number of variables. These coefficients are called *loadings*. J loadings form a *loading vector* **p**, which is a linear combination of vectors defined by the measured parameters in the multivariate direction that span the most variance. This is the same as minimizing the least squares distance from all the samples to the loading vector that eventually will form part of the PC. When the first principal component (PC#1) is identified, all samples will be projected orthogonally on to this loading vector. The information about the position of the individual samples along this axis is called the *score* of the sample with respect to the PC. Information about all K samples can be compiled into a *score vector* **t**. After the first PC is identified, the information about the direction of the first PC and the scores of the samples are subtracted from **X**. Subsequently, the second principal component (PC#2) is identified. As the information from PC#1 has been removed the direction of PC#2 will be orthogonal to the direction of PC#1. The third principal component (PC#3) will be orthogonal to both PC#1 and PC#2, etc. PC's can continue to be extracted until all the information in **X** is completely described. This will happen at the mathematical rank – which was the minimum of K or J. In that case, all the information in **X** is rotated from the original variables to a new set of variables defined by the loading vectors (which are linear combinations of the original variables), which are orthogonal to each other and sorted according to the variance in **X** they describe. **X** can now be split in a structural part with only the PC's describing "large" or systematic variation, and a noise part with the PC's responsible for "small" variation, less systematic and more random. If it is later chosen to describe **X** with just the PC's responsible for the structural variation, the dimensionality – and the noise – of **X** can be reduced, while the main variation of **X** is preserved. The benefit may not seem obvious if only a few variables are measured, but in the multivariate situation, a reduction from several thousand variables to a three or seven dimensional subspace, the benefit can be immense.

All loadings combined in one principal component form a loading vector, **p**. Several loading vectors can be combined to form a loading matrix, **P**. A loading plot is one or more loading vectors plotted against each other with the loadings on the axis, or the original variables on one of the axis, usually the x-axis. A loading plot is a map of the original variables and represents by

which weight the original variables form part of the principal components. Scores are the orthogonal projection of the objects on to the individual loading vectors. The scores are the coordinates of the axis as defined by the loadings. All scores from one principal component combined form a *score vector*, **t**. They can be calculated as the inner product of the vector defining the objects position in the original variable space and the loading vectors. All score vectors are orthogonal to each other and can be combined to form the score matrix, **T**. A score plot is a map of the objects' scores usually plotted against each other in a scatter or (pseudo) three dimensional plot. A score plot can be interpreted as a map of the objects (Esbensen 2000). By convention, the length of the loading vectors is set to one, i.e. they are ortho-normal. The size of each PC is preserved in the length of the score vectors to eliminate the scaling ambiguity.

For computational purposes the measurements are often centred. Centring can lead to reduced effective rank of the model, increased fit, remove irrelevant offsets and simplify the calculations. Centring two-way matrices is comparable to removing the levels of the variables from the elements in **X** so that the average of the columns in **X** is 0:

$$\mathbf{X}_{jk} = \mathbf{X}_{jk} - \Sigma_k(\mathbf{x}_{jk})/K \qquad (5)$$

Where the last term is the average of column *j* in **X**. One offset is subtracted from every element in each variable column. Centring of multi-way matrices is not so trivial, but two-way and multi-way centring can be generalized to a projection step, where the data are projected onto a chosen sub-space within a given mode (Bro, 2003). Figure 16 shows the effect of centring on the spectroscopic data as presented in **Paper I.**

The data used is the "Synthetic" samples and the "Process calibration set" – see the paper for full details. On the left hand side of the figure, a spectroscopic offset in the measurements from "Tank 1" is identified. This is due to an unintended larger path length in the sample cell measuring amidation liquid from Tank 1. As seen on the right hand side of the figure centring does not remove the offset. Centring allows PCA to use the first principal component to actually describe the variation between the samples rather than the variation of the samples to an imaginary "no absorbance" sample at the centre of gravity of the original variable space).

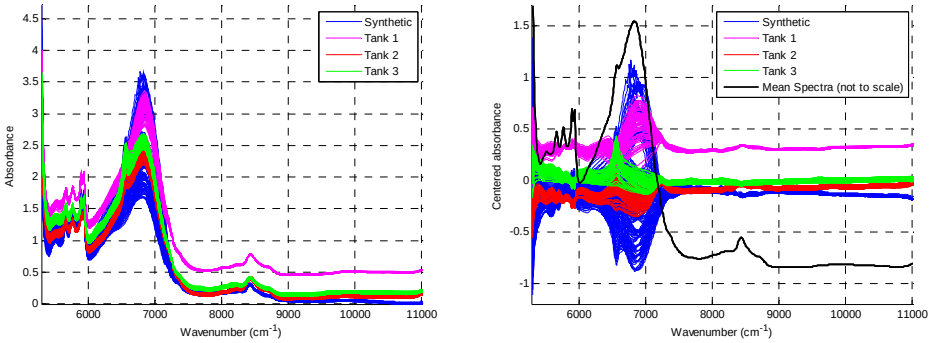


Figure 16 Spectra of amidation liquid from the calibration set presented in **Paper I**. To the left are the raw spectra and to the right are the centred spectra with the mean spectra removed in the centring calculations overlaid.

In matrix notation, the PCA model can be written as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (6a)$$

X is the raw data (of size $K \times J$), **T** is the score matrix with the same number of rows (K) as **X**, **P** is the loading matrix with the same number of rows (J) as variables as **X**. **E** (also of size $K \times J$) is the residual matrix holding whatever is left unexplained by the model which is the matrix product of $\mathbf{TP}^{\mathrm{T}}$.

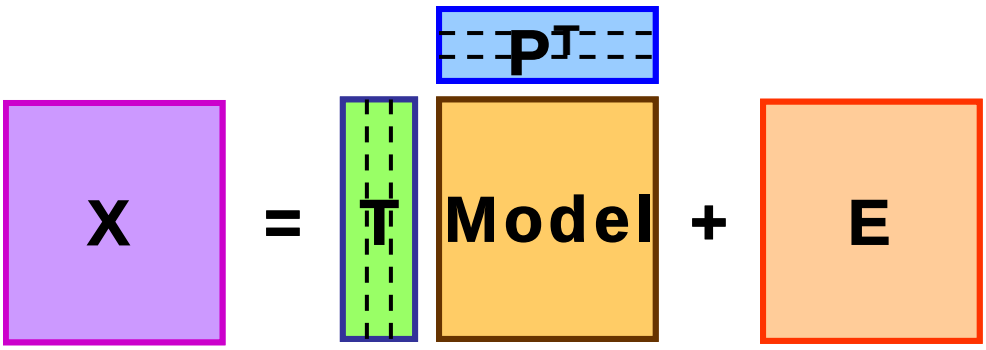


Figure 17 Illustration of a PCA model.

If (6a) is written using individual vectors for $\mathbf{TP}^{\mathrm{T}}$, then when F principal components are extracted:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_F\mathbf{p}_F^{\mathrm{T}} + \mathbf{E}_F \qquad (6b)$$

Figure 18 illustrates the PCA model vector-wise i.e. with the individual principal components. The outer product of $\mathbf{t}_1$ and $\mathbf{p}_1^{\mathrm{T}}$ gives the first principal component's contribution to **X**, the outer product of $\mathbf{t}_2$ and $\mathbf{p}_2^{\mathrm{T}}$ gives the

second principal component's contribution to **X** and so on. $\mathbf{E}_F$ is the residual matrix (of the same size as **X**) when F components are extracted.

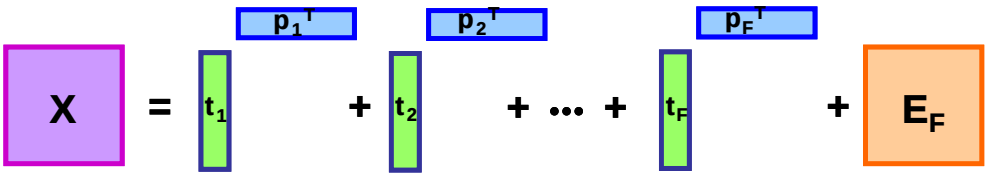


Figure 18 Graphical illustration of a PCA model. $\mathbf{t}_1$ and $\mathbf{p}_1$ are the score and loading vector to the first principal component respectively.

Even though the residuals are not part of the PCA-model, the analysis of the residuals is as important as the analysis of the model itself. The optimal number of principal components to extract is determined by analyzing the residual matrix after extraction of each principal component; that is $\mathbf{E}_1$, $\mathbf{E}_2$, … in principle up to the full rank of **X**. The squared sum of all elements in **E** can be compared with the squared sum of all elements in **X**. When **X** is fully described by the PCA model, **E** will be a matrix of zeros. A plot of the squared sum of all elements in **E** as a function of the number of principal components extracted will disclose if there is an effective rank of **X** lower than the mathematical rank. Residual analysis on rows and columns of **E** also indicate whether all objects or variables are described well by the model or if outliers in the objects exist or if some variables are poorly explained by the model.

An example of a PCA

Figure 19 shows the explained variance in percent of a PCA on the full spectral region of the "Synthetic" samples and the "Process calibration set" from **Paper I**. It is a PCA on the spectra presented in Figure 16. The first four principal components explain 85.9%, 9.8%, 2.6% and 1.1%, respectively. The information from the 402 samples and 740 wavelengths can be compressed to four principal components, while retaining 99.4% of the original variance in the data, compared to the centred NIR spectra. Eight principal components account for 99.99% of the total variance.

The loading plot can be seen as a map of the variables. A loading plot of the PCA can be seen in Figure 20. It reveals which of the originally measured variables are important for the principal components.

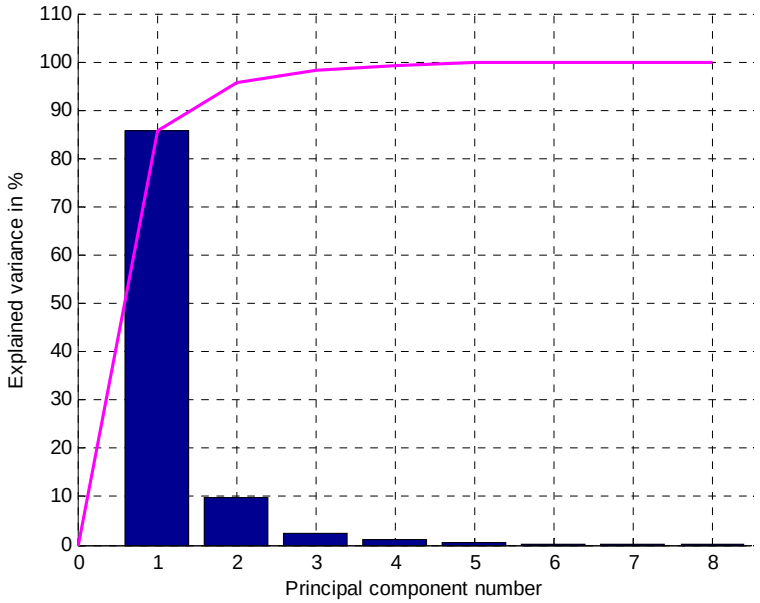


Figure 19 The explained variance in percent as a function number of principal components from 1 to 8 extracted from the calibration set from **Paper I**. The bars are the explained variance by each component, the curve is the cumulative variance explained in percent.

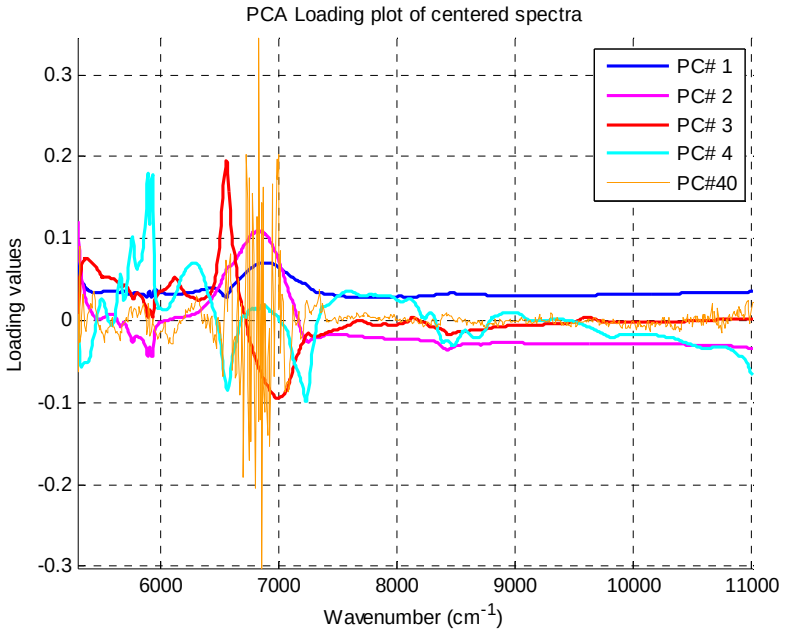


Figure 20 Loading plot of PC#1 to PC#4 and PC#40. Even though the 40th PC appears more noisy than the first four PC's, some degree of structure is still evident.

Several distinct features from the mean spectra as seen on the right side on Figure 16 can be identified. These can be related to spectral fingerprints of water (PC#2), ammonia (PC#3) and Isopropyl alcohol; IPA (PC#4). See also the following chapter on pectin; especially Figure 46. Even though the first four PC's capture 99.4% of the total variance, there is still a lot of structure left in the residuals. As an example, PC#40 is also shown in the plot. PC#40 accounts for only 0.000011% of the total variance and the first 40 PC's explain 99.9998% of the variance in total. Albeit noisy, it is far from white (random) noise. Some spectral features, especially between 5300 and 6500 cm^{-1}, as well as proportional noise are present.

A score plot can be seen as a map of the samples. Various score plots as combinations of PC#1 to PC#4 can be seen in Figure 21. From the appearance of the loading plot combined with knowledge of the samples, the sample acquisition, NIR spectroscopy and other available information, an interpretation of the principal components can be made. PC#1, the most dominant variation by far is associated primarily with the absorbance level or baseline of the spectra. As seen in Figure 16, measurements from Tank 1 have a higher absorbance than the other samples. This is probably because the measurement cell from Tank 1 inadvertently has a longer path length than the other measurement cells. PC#1 therefore reflects the physical conditions of the measurements rather than the sample itself. PC#2 has positive loadings in the water region and negative loading in the IPA region. The amounts of water and IPA are inversely correlated. PC#3 reflects ammonia and PC#4 has information associated with all three components of IPA, ammonia and water. Not surprisingly measurements from Tank 1 form a separate cluster along PC#1. PC#4 separates measurements from Tank 2 from the other measurements, while PC#3 vs. PC#4 show a reasonable distribution of both the Synthetic (laboratory prepared) samples as well as the measurements from the tanks. It compares relatively well with the PLS score plot on Figure 9 in **Paper I**. The residuals, the part which is not modelled by the principal components, often contain as much information as the model itself. The sum of the squared sample residuals can be seen on the left part of Figure 22, and the same sum the other way in the residual matrix gives the variable residuals on the right. The sample residuals reveal that the first 152 samples which happen to be the Synthetic samples have high residuals. There are still unexplained features left. Measurements from Tank 1 on the other hand, have low residuals and are therefore well explained by the four PC's. The sample residual plot is important for the identification of outliers; samples which are not modelled well and possibly do not belong to the

same population as the rest of the samples in the study. The variable residuals reveal that the shoulder right at 5300 cm^{-1} is poorly explained by the first four PC's as well as part of the water peak from 6600 to 7200 cm^{-1}.

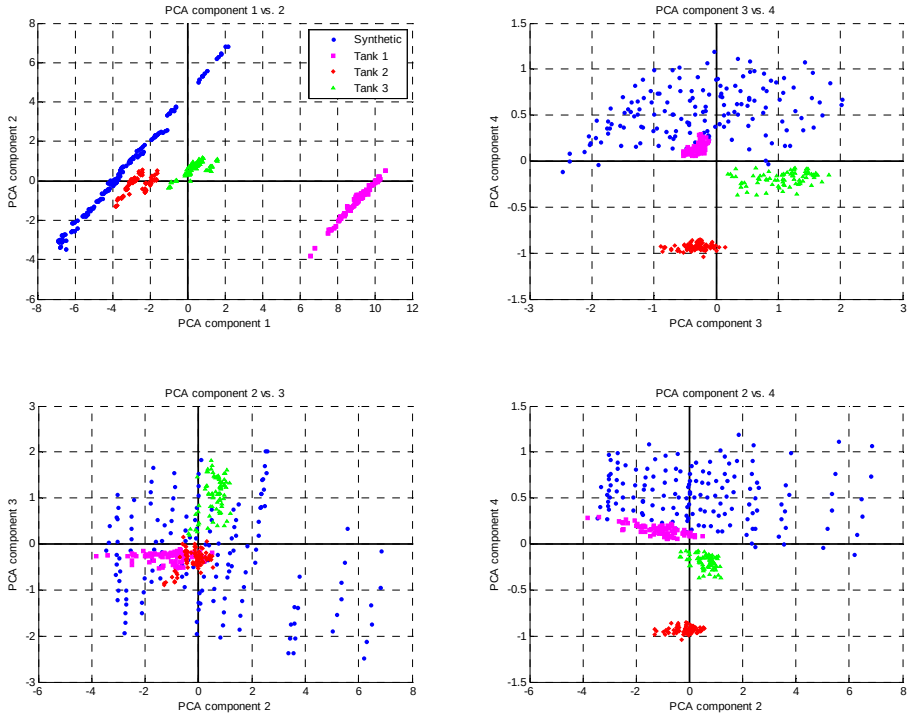


Figure 21 Different score plots can be achieved by combining the score values from the first four principal components. The first four principal components explain 85.9%, 9.8%, 2.6% and 1.1% of the variance, respectively.

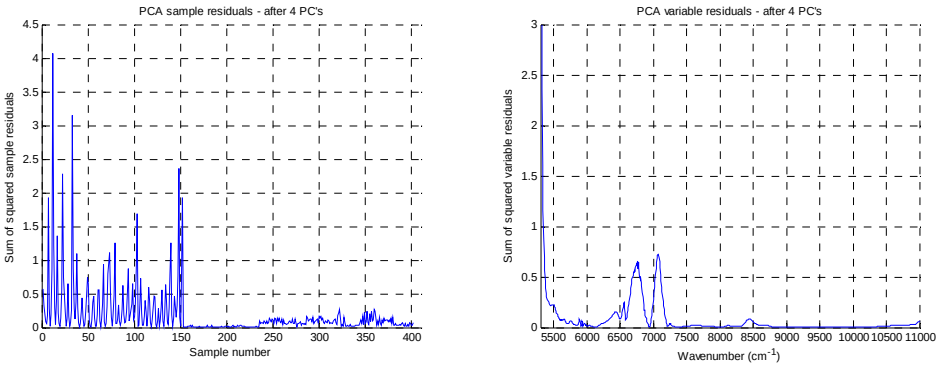


Figure 22 The sum of the squared sample residuals (left) and the variable residuals (right) after four principal components are subtracted. The sample/variable residuals are a measure of how much of each sample/variable remains to be accounted for by the model.

Multivariate Curve Resolution (MCR)

Multivariate curve resolution (MCR), see Figure 23, is a commonly used technique that can resolve multi-component mixtures into a simple model consisting of a composition-weighted sum of the signals of the pure compounds (Tauler, 1995a; Tauler, 1995b; de Juan, 2003). As such, MCR analyses *one* sample characterized by second order data. The multivariate curve resolution model can be written as:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \qquad (7)$$

It is a bilinear method to resolve an experimental data matrix **D** ($I \times J$) into the product of a column matrix **C** ($I \times F$) usually associated with concentration profiles, and a matrix of row profiles $\mathbf{S}^T$ ($F \times J$), usually associated with spectra. The matrix **E** ($I \times J$) is the residuals i.e. what is not explained by the model $\mathbf{C}\mathbf{S}^T$. The entries in matrix **E** should be small and random compared to the numbers in **D** and the model $\mathbf{C}\mathbf{S}^T$. The scalars I and J are the total number of time points in the data set studied and wavelengths of the data set, respectively, and F is number of components to be resolved. This general description can fit a wide range of applications, processes, mixtures, elutions, images and environmental data monitored by multivariate instrumentation (Jaumot, 2005; de Juan, 2003). The model parameters are estimated using an Alternating Least Squares (ALS) algorithm that iteratively fits the **C** and $\mathbf{S}^T$ matrices to the experimental data **D**. The model is fitted with a pre-defined number of components, F, using initial estimates of either the **C** or the $\mathbf{S}^T$ matrix.

Exploratory data analysis using Singular Value Decomposition/PCA, Evolving Factor Analysis (EFA; Maeder, 1986; Maeder, 1987; Keller, 1992) or Simple-To-Use Interactive Self-Modeling Mixture Analysis (SIMPLISMA; Windig, 1991) can provide knowledge used for the initial estimates. The MCR algorithm and constraints that can be applied during iterations is further discussed in **Paper II**.

A powerful extension of MCR is multi-sample MCR, where it is possible to resolve several independent samples and/or several independent measurement techniques by augmenting the input matrix appropriately (Tauler, 1995a; Tauler, 1995b; Jaumot, 2005). Extending the concentration matrix column-wise to estimate concentration profiles of several samples will facilitate the resolution of a global solution. This is taken advantage of in **Papers II and III**.

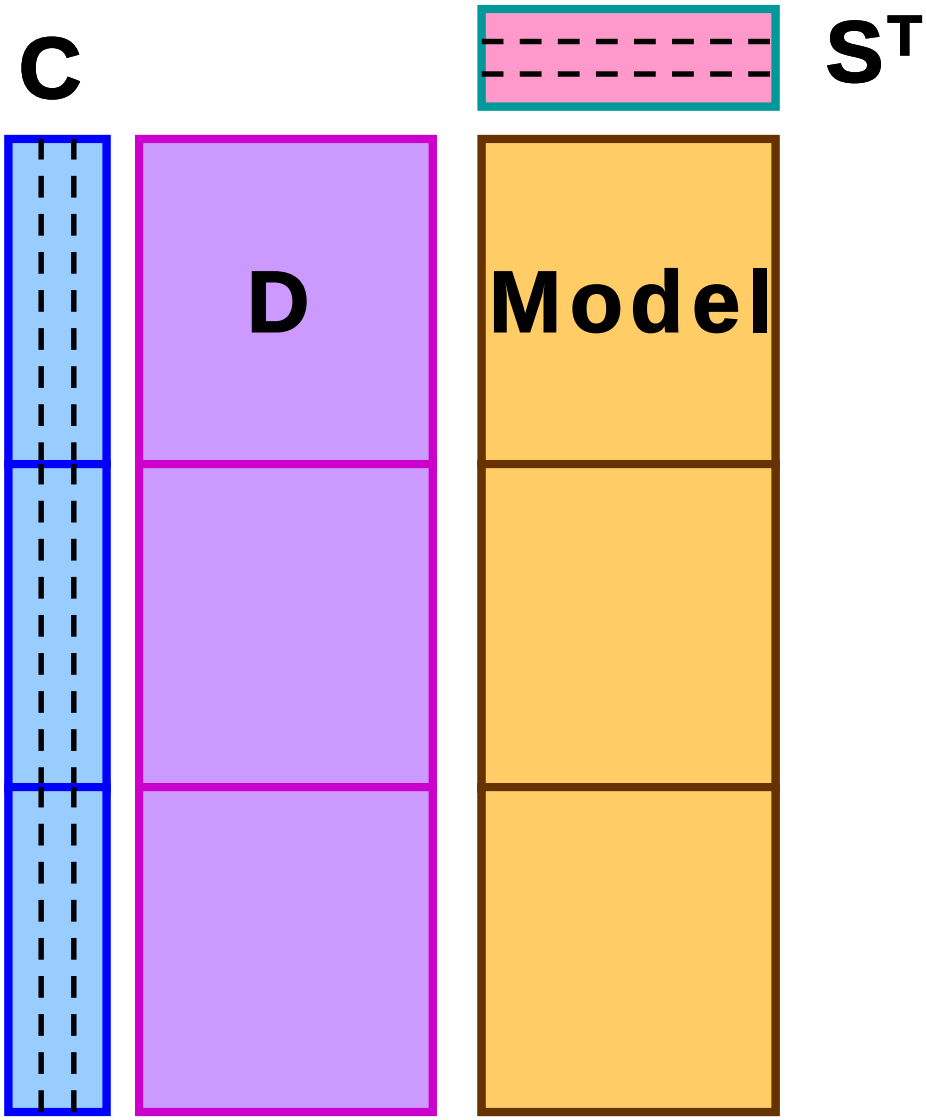


Figure 23 The Multivariate Curve Resolution model extended to several samples and assuming that the same spectra are present in all the samples, but with different concentration profiles.

Parallel Factor Analysis (PARAFAC)

PARAFAC is an acronym of PARallel FACtors. The algorithm was simultaneously developed by Harshman (Harshman, 1970) and Carroll and Chang (Carroll, 1970) and originally developed for psychometrics, the measurement of "psychological" aspects of a person such as knowledge, skills, abilities, or personality. PARAFAC has later been applied to chemometrics (Ge-

ladi, 1987; Bro, 1997). The PARAFAC algorithm can be seen as a multi-way extension of PCA. In the three-way case, the algorithm models a three-dimensional array **X**. The elements of **X** can be computed the following way by PARAFAC:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad (8)$$

where x_{ijk} represents an element in **X**, in the position given by i, j and k. a, b and c are the loadings and e_{ikj} is the residual; the unmodelled part of the data. The rank of the PARAFAC model is given by the number of factors, F, needed to describe the variation in the data array. Vector-wise a three-way PARAFAC can be seen as a summation over outer products of F triads of vectors (Figure 24), and the similarity with PCA is obvious. In PARAFAC, the term scores are some times still used for the loadings in sample mode but most often just loadings are used irrespectively of which mode is referred to. In PARAFAC it is necessary to compute all factors simultaneously, and because of the second-order advantage it will give unique solutions. This makes PARAFAC an excellent tool for curve resolution (Bro, 1997). PARAFAC, as it is, is however not able to handle time or retention shifts and is not used in **Papers II and III**.

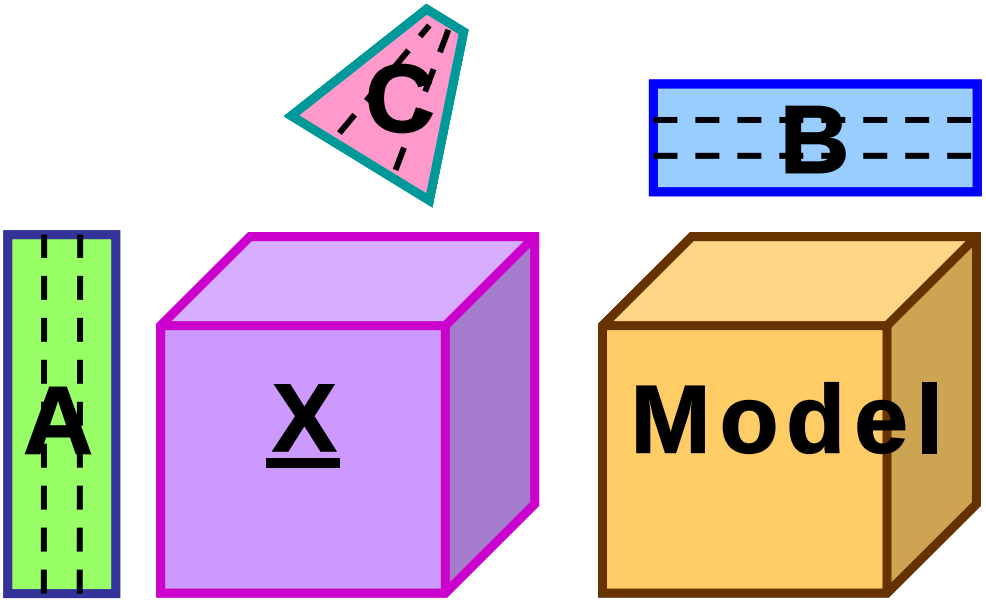


Figure 24 The PARAFAC model. The triads of vectors are here combined to loading matrices **A**, **B** and **C**.

PARAFAC2

PARAFAC2 (Figure 25) is an advanced variant of the PARAFAC algorithm developed to handle the situation where the number of observations in one mode varies or where shifts or shape-changes of profiles along one mode are anticipated. This makes the algorithm appropriate to handle, for instance, retention time shifts (Bro, 1999). The scalars I, J and F hold the number of time points and wavelengths of the data respectively, and the number of components to resolve. The scalar K holds the number of samples. The matricized notation of the PARAFAC2 model can be written as:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k(\mathbf{P}_k\mathbf{H})^T + \mathbf{E}_k, \quad k = 1, .., K \qquad (9)$$

Where $\mathbf{X}_k$ represents the data related to one sample (one slab from the original data cube) of size ($J \times I$) in which I can vary with k. K is the number of samples. $\mathbf{A}$ ($J \times F$) holds the first-mode loadings, in this study the resolved spectra. $\mathbf{D}_k$ ($F \times F$) is a diagonal matrix that holds the k'th row of $\mathbf{C}_{P2}$ in its diagonal. The notation $\mathbf{D}_k$ is maintained for consistency with literature, but note that $\mathbf{D}_k$ is not related to the $\mathbf{D}_k$ used in the MCR model. $\mathbf{C}_{P2}$ ($K \times F$) holds the third mode loadings, usually the sample scores. $\mathbf{H}$ is an $F \times F$ scaling matrix, and $\mathbf{P}_k$ is an $I \times F$ ortho-normal matrix (I can vary with k). The matrix $\mathbf{E}_k$ holds the residuals. $\mathbf{P}_k$ and $\mathbf{H}$ have no direct chemical or physical interpretation but their product will be an estimate of the time profiles.

The PARAFAC2 model differs from a Principal Component Analysis (PCA) of the unfolded array $\underline{\mathbf{X}}$ by the requirement that the spectral loadings $\mathbf{A}\mathbf{D}_k$ must be proportional for every value of k (running from 1 to K). Furthermore, it is assumed that the cross-product $(\mathbf{P}_k\mathbf{H})^T(\mathbf{P}_k\mathbf{H})$ is constant for all values of k (Kiers, 1999; Wise, 2001). This condition is sufficient for the PARAFAC2 model to be unique if some other mild uniqueness conditions are met, as further discussed in depth in Kiers et al. (Kiers, 1999). Even though the matricized notation of PARAFAC2 is presented, it should still be considered a multi-way method with an independent sample mode. In MCR-ALS, the sample mode is only implied through extension of the concentration matrix.

The PARAFAC2 model is more flexible, compared to the PARAFAC model. It allows for deviations of the inherent trilinearity imposed by the PARAFAC model. The concentration profiles resolved on the time axis are allowed to vary as long as the before mentioned cross-product is constant over different samples. More about PARAFAC2 as well as the similarities and intricate differences between PARAFAC2 and MCR are discussed in **Paper II** and (Smilde, 2004) and in references therein.

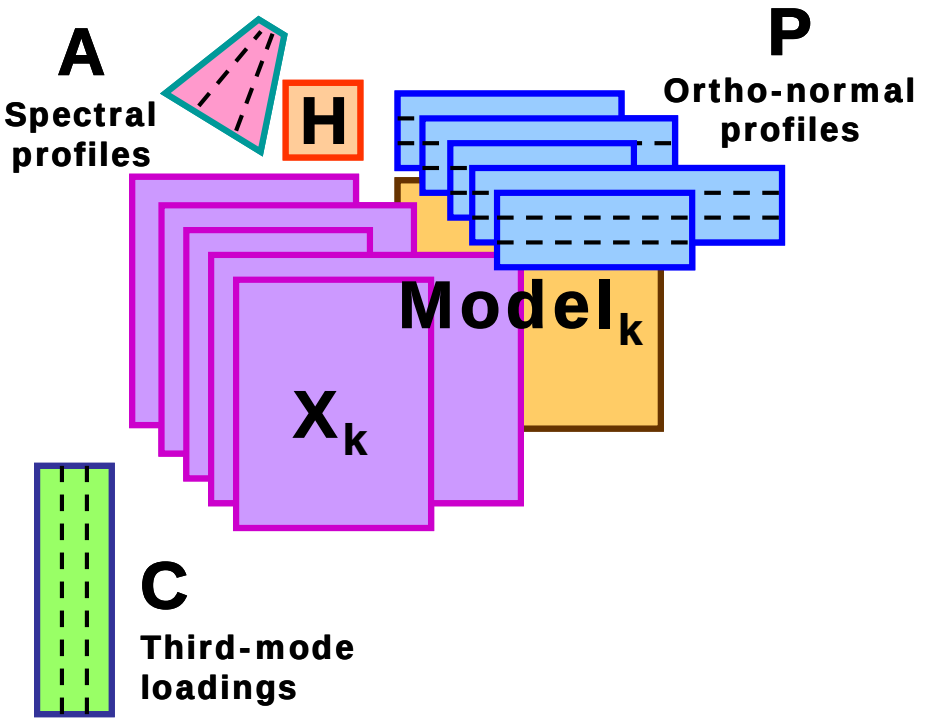


Figure 25 The PARAFAC2 model. The advantage of PARAFAC2 over PARAFAC is that it can be applied on data of unequal length in one of the dimensions, and that it can handle time shifts. Only one model matrix is shown.

Partial Least Squares (PLS) Regression

The purpose of Partial Least Squares regression (PLS) is to model a set of dependent variable(s) **y** or **Y**, from a set of independent variables **X**, characterised in the same manner that was valid for the description in the section of PCA. The purpose of this is to estimate **Y** from future measurements of **X** only. This requires a *calibration set* with known values of **X** and **Y**. The calibration set should be representative (Gy, 1998; Gerlach, 2003; Petersen, 2005) for future populations of **X** and **Y** with regard to known and foreseen variations, which should also be spanned as well as possible.

PLS is a commonly used and well described method in chemometrics (Sjöström, 1983; Esbensen, 2000; Wold, 2001). In PLS there are two simultaneous decompositions, see Figure 26. As in PCA the **X** matrix is decomposed into **T** and **P**, the score and loading matrices respectively. **Y** is similarly decomposed into **U** and **Q**, which are the score and loading matrices in the **Y** space, respectively. The difference between PLS and merely two PCA decompositions of the **X** and **Y** blocks is that the decomposition of **Y** influences

the decomposition of **X** and the opposite way around. This is done by estimating PLS components that capture and maximize the variance and correlation between **X** and **Y**.

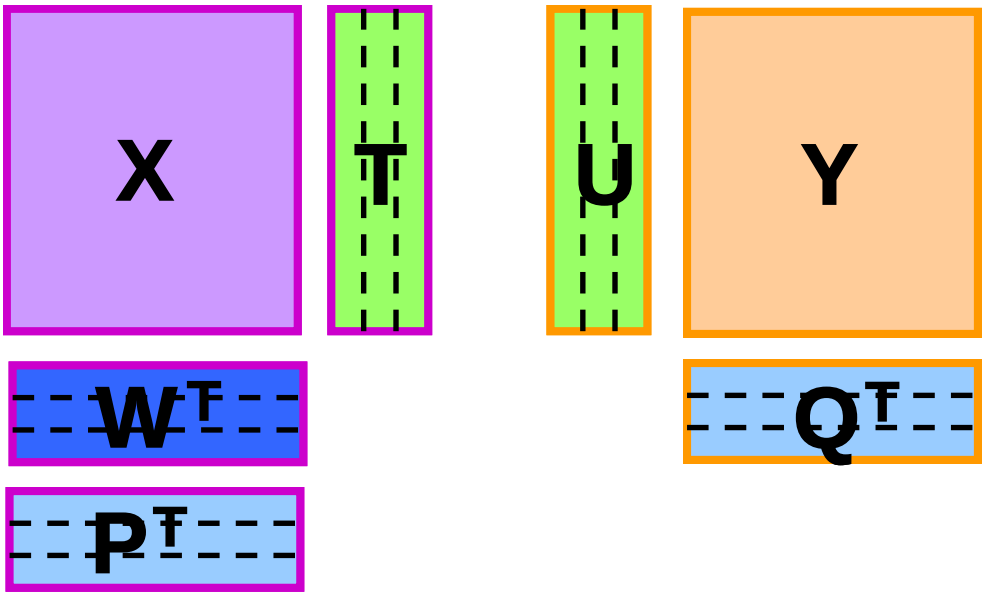


Figure 26 The matrices involved in the PLS-algorithm. **T** and **P** together with **U** and **Q** are score and loading matrices for the **X** and **Y** block, respectively. **W** are the loading weights that express which variables in **X** are important for the description of **Y**. The residual matrices are not shown.

The PLS model can be regarded as consisting of an *outer relation* and an *inner relation*, where the outer relation describes the **X** and **Y** block, while the inner relation links the two blocks together. The outer relations are given by:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_X \quad (10)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{E}_Y \quad (11)$$

Where $\mathbf{E}_X$ and $\mathbf{E}_Y$ are the residual matrices for **X** and **Y**, respectively.
The inner relations are given by:

$$\mathbf{U} = \mathbf{B}_D\mathbf{T} + \mathbf{H} \quad (12)$$

where $\mathbf{B}_D$ is a diagonal matrix with the regression coefficients from the inner relation and **H** is the inner relation residual matrix. The loading weights, **W**, can be computed from the estimation of the scores, T, in the *X*-block:

$$\mathbf{W} = \mathbf{T}^T\mathbf{X} \quad (13)$$

The loading weights, **W**, are related to the PLS regression coefficients **B** by:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T \qquad (14)$$

The regression coefficients **B** can be used to predict new **Y** values from new **X** values:

$$\mathbf{Y}_{new} = \mathbf{X}_{new}\mathbf{B} + \mathbf{E}_{Y,new} \qquad (15)$$

For further details of the PLS algorithm, the reader is referred to (Haaland, 1988; Bro, 1996b; Wold, 2001; Smilde, 2004) and references therein. PLS has successfully been applied to higher order data both of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$. This is called multi-way Partial Least Squares regression (or simply N-way PLS or N-PLS; Bro, 1996a; Smilde, 2004).

Validation and error measures

An important element in chemometric data analysis is the validation of the calculated models. This is to avoid false/chance correlations, determine the optimal number of PCA or PLS components to use in the final model and to ensure that the estimated model reflects the reality. If not, no conclusions should be drawn from the model. The best validation is a completely independent set of new objects or samples – a so-called test set – that *if* it reflects the variation in the calibration set *and* the future samples, will give a realistic estimation of the error. This is also referred to as *external validation*. If an independent test set is not available, it could be defined from the available objects. This should be done by spanning the variation in population as well as in time. An independent test set is used after the final model is computed for assessment or calculation of e.g. the prediction error. If an "independent" test set is used in phases of the modelling, it is no longer independent.

If no independent test set is available, for instance if the number of samples is limited and/or during the model development using the calibration set, some sort of *internal validation* is needed. One example is *cross validation*. Predefined segments of the sample set are excluded in the modelling phase, and the model is calculated on the remaining data. The excluded data are then fitted and/or predicted on the model developed on the segmented data. Other data are then in turn left out until all predefined segments have been kept out. For the final model all samples are used, but the error measures computed during the cross validation step is used for performance evaluation. In PCA, the size of the sample residuals can be used as a measurement of the modelling error, and in PLS, the prediction error of the **Y** data is used. The recommendations how to divide a sample set into test set and calibra-

tion set as well as when to use cross validation and test set validation, and *if* cross validation is to be used, then how to define the segments, are much debated in scientific literature (Wold, 1978; Martens, 1989; Martens, 1998; Louwerse, 1999; Esbensen, 2000; Smilde, 2004). It will always be related to the data set at hand and the purpose of the modelling.
A commonly used way of estimating the prediction error is in terms of the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{i,p} - y_{i,r})^2}{n}} \qquad (16)$$

where $y_{i,p}$ for sample number i is the predicted value, $y_{i,r}$ the actual reference value for sample number i, and n the total number of samples. The RMSE-values are in the same units and scale as the reference values. Depending on how the model is used for estimating the predicted values, terms like RMSEC (Root Mean Square Error of Calibration), RMSECV (Root Mean Square Error of Cross Validation) and RMSEP (Root Mean Square Error of Prediction) are used. Occasionally RMSEP is used interchangeably with RMSECV, but RMSEP should only be used with test set validation.

Pre-processing of spectra
As seen in Figure 16 and Figure 21 baseline shifts, scatter effects and other physical phenomena may interfere with the chemical information recorded in the spectra. Spectral pre-processing may be used to correct for these effects, and improve performance and interpretation of calibration models. Many such methods exist, and in the following section some of the most common will be presented. Multiplicative Signal Correction (MSC; Geladi, 1985) corrects for additive (a level correction) and for multiplicative differences in the spectra. The method was originally developed to the correction of scatter effects (and is hence also known as Multiplicative Scatter Correction) that is commonly seen in spectral measurements of powders, particles of different size distribution and slurries, but has also proved effective in correcting for instrument and temperature drift and other physical interferences. MSC could be done on the entire spectra, but often selective regions without absorbing species should be selected as a basis for estimating the additive and multiplicative effects (Esbensen, 2000). One disadvantage with MSC is that the mean spectrum has to be stored for future MSC transformations, and if suddenly new spectra changes significantly, the MSC will not handle them in a proper way. Recently the method has been further devel-

oped to Extended Multiplicative Signal Correction (EMSC; Martens, 2003; Thennadil, 2006) which corrects for quadratic effects as well. The Standard Normal Variate (SNV) pre-processing is theoretically related to MSC (Dhanoa, 1994). The SNV pre-processing is done for each individual wavelength in each individual spectrum and is not dependent on a global mean spectrum.

A different approach to eliminating the additive and multiplicative signal effects in the spectra is by computing the derivatives. The first derivative will remove additive effects and taking the second derivative will remove both additive and multiplicative effects. In order not to amplify noise in the spectra a smoothing filter is often applied simultaneously. The most used method for computing derivatives is the one proposed by Savitzky and Golay (Savitzky, 1964). The algorithm fits a polynomial for each individual data point in a pre-defined window size symmetric left and right of the data point of interest. Thus, some parameters have to be decided upon before applying the algorithm. Obviously, taking the derivatives changes the appearance of the spectra, and complicates spectral interpretation.

Figure 27 shows the effect of the above mentioned spectral pre-processing methods on the NIR spectra from **Paper I**. Only the most interesting spectral region from 5300 to 9000 cm^{-1} is shown. That pre-processing can assist in removing unwanted information can also be seen by relating the individual wavelengths of the spectra in a calibration set to the property of interest. In Figure 28 it is clearly seen how the direct correlations between the individual wavelengths and the property of interest improve by pre-processing the data. The data used are the calibration set from in **Paper I**, and the property of interest is ammonia. There are also pre-processing methods that actively use information about the property of interest. Orthogonal Signal Correction (OSC; Wold, 1998), Direct Orthogonalization (Andersson, 1999) and Net Analyte Signal Correction (Lorber, 1997) all use some PLS-like variants in a pre-processing step to remove unwanted variation from **X** that does not correlate to **Y**. The effect of various pre-processing methods on the prediction of ammonia on the data set of **Paper I** can be seen in Figure 29. The various models are computed using the calibration set and applied on the test set. This is only done for illustrative purposes for this thesis. To make plots like these during the modelling stage would be improper use of the test set. The performance improvement by pre-processing can clearly be seen as the Raw (not pre-processed) prediction error is the highest in most instances. Note that the EMSC pre-processing leads to a dramatic increase in prediction error when seven PLS components are used in the full spectrum model.

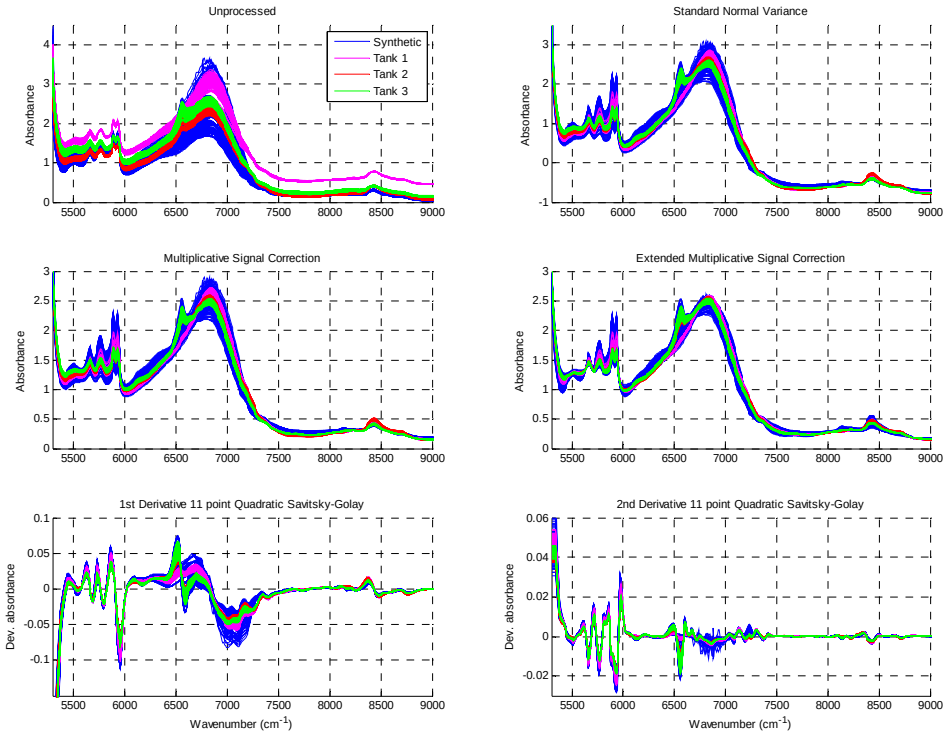


Figure 27 The effect of different pre-processing methods on NIR spectra. Second degree polynomials are fitted to the individual data points using a window size of 11 (5 points to the left and right of the data point in question) to compute the derivatives.

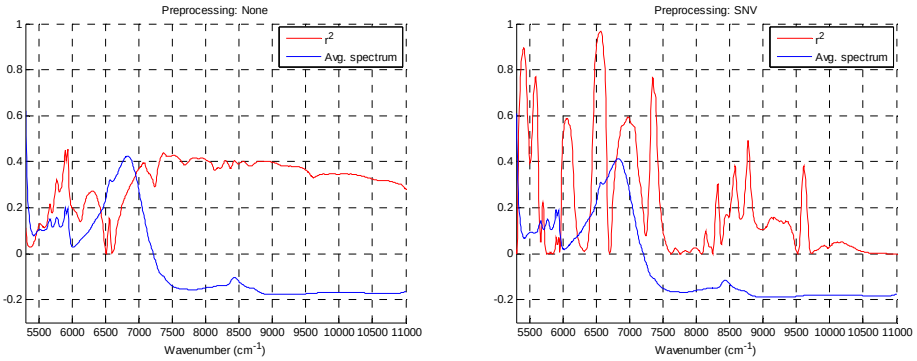


Figure 28 The squared of the correlation coefficient (r^2) of the NIR spectra in the calibration set to the ammonia content. When no pre-processing is done the r^2 is below 0.5 at all times while it is close to 1.0 in certain wavelength regions when the spectra have been pre-processed with SNV. The average spectra in the figures are not to scale.

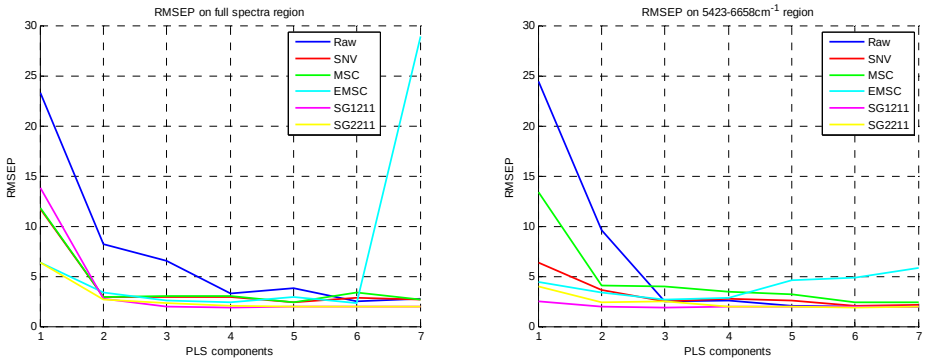


Figure 29 The effect on various pre-processing methods on the prediction of ammonia. The prediction error (RMSEP) estimated using test set validation is plotted as a function of the pre-processing method and the number of PLS components used. To the left is models calculated using the full spectral region, on the right is only the spectral region of 5423 to 6658 cm^{-1} used.

For reasons of confidentiality, the concentrations of IPA and ammonia have been rescaled to 0-100% following the convention used in **Paper I**.

Variable selection

As can be gathered from the right hand side of Figure 28, and what is known from the theory of spectroscopy, certain wavelength regions contain specific information about the analyte of interest. In Figure 30 this is further illustrated by colouring the spectra in a scatter plot according to the amount present of the analyte of interest.

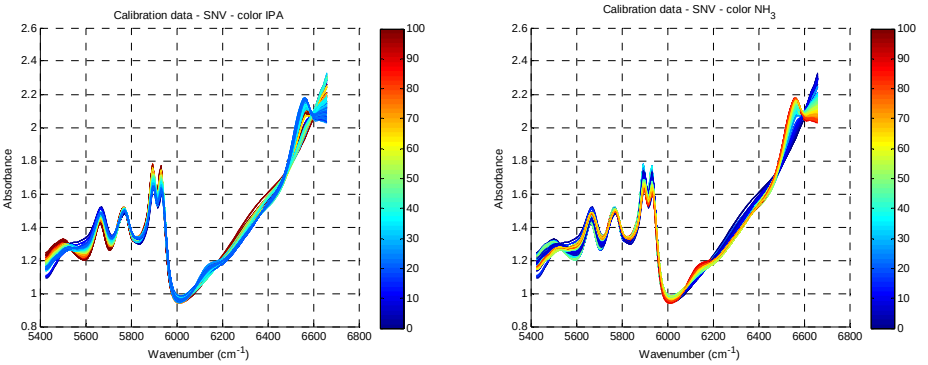


Figure 30 The NIR data from the calibration set of **Paper I** is pre-processed using SNV and the individual NIR spectra are coloured by the content of Isopropyl alcohol (IPA, left) or ammonia (NH_3, right). The region from 5423-6650 cm^{-1} is plotted. Several peaks which correlate to the analytes can be identified.

It is seen that calibration performance can be improved by eliminating regions with no relevant information. Several approaches can be taken to variable selection, for instance genetic algorithms (Leardi, 2001) or interval PLS (iPLS; Nørgaard, 2000). A simple, but efficient way is to plot all combinations of one-region selections of the wavenumbers. In **Paper I** however, 740 wavenumbers are recorded in the region of 5300 to 11001 cm^{-1}. 274,170 different regions can be picked, but by taking the iPLS approach and combining *intervals* of wavenumbers, the number of models to compute can be reduced significantly. By combining 15 wavenumbers equivalent to about 115 cm^{-1} intervals, the number of combinations of regions is lowered to 1225. In Figure 31 an "error surface" for *all* combinations of *continuous* intervals giving RMSEC (using the calibration set) and RMSEP (estimated using the test set) values is compiled for each PLS component and represented as a coloured pixel. The spectra are pre-processed by taking the first derivative using a quadratic Stavitzky-Golay algorithm using a window size of 11 points, and the property of interest is ammonia. In each of the plots the *start* wavenumber is indicated on the y-axis and the *end* wavenumber is on the x-axis. In the lower left corner of each figure is the first row of (in this case 49) pixels representing the models computed starting at 5300 cm^{-1}. In the lower right corner is the full spectrum model starting at 5300 cm^{-1} and ending at 11000 cm^{-1}. In the row above are the 48 pixels representing the error of models starting at ~5415 cm^{-1}. In the top right corner is the sole pixel representing the model starting at ~10820 cm^{-1} and ending at 11000 cm^{-1}.
The regions with low prediction error are the blue areas, and in general the more PLS components that are used the lower is the prediction error. If the RMSEC error surface using 2 PLS components (placed upper-right) are scrutinized several interesting regions can be identified, notably the region starting from 5400 to 6000 cm^{-1} and ending at 7000 cm^{-1}. But other interesting regions with low RMSEC is the region starting from 7500 to 7700 cm^{-1} and ending at 11000 cm^{-1}. The red regions have high RMSEC and do not hold information of the ammonia content.

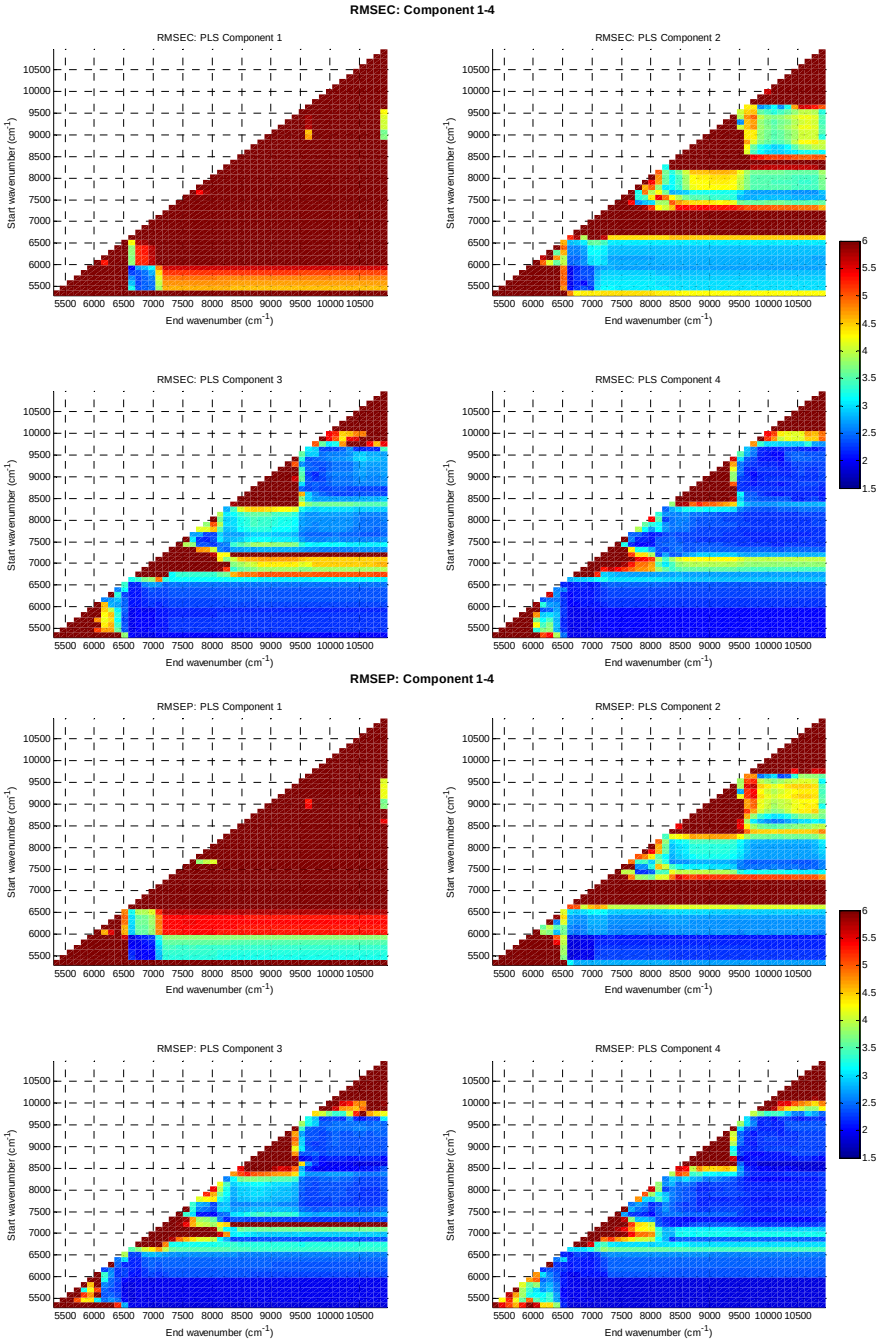


Figure 31 RMSEC and RMSEP computed and its values are shown as coloured pixels for 1225 PLS models for one to four PLS components. Blue areas indicate wavelength combinations with good calibration performance (low error). See the text for further details.

4. Pectin

An industrial commodity

Pectin is a polysaccharide found in the primary cell walls and intracellular regions of all higher order plants. Pectin was first isolated and described in 1825 by the French chemist and pharmacist Henri Braconnot. He also suggested the French name *pectine* based on the Latin word *pectis*, which means coagulated. The original Greek word *pēktos* is derived from *pēgnunai*, to coagulate. Pectin has excellent gelling, thickening, texturizing and stabilizing properties and is widely used for commercial applications, primarily in the food industry. Commercial production of pectin is done using mainly citrus peel and apple pomace available as a by-product from industrial juice and cider manufacturing (Rolin, 1990; FAO, 2001; Rolin, 2002). Annually 4000 million litres of orange juice and 450 million litres of grape fruit juice are produced (leaving theoretically more than 600 million tons of dried peel). Furthermore, 5 million tons of apples are processed into juice every year (Madden, 2000). Apple pomace contain about 10-15% and citrus peel about 20-30% pectin on a dry weight basis. A small amount of commercial pectin is produced from sugar beet pulp, primarily used to stabilise emulsions. Pectin has always been part of human foods. As pectin is considered safe, the Acceptable Daily Intake (ADI) of pectin is "not specified" by the Joint FAO/WHO Expert Committee on Food Additives (JECFA) as well as in the EU, and the US Food and Drug Administration renders pectin the Generally Recognized as Safe (GRAS) status. In the International Numbering System (INS) pectin has the number 440. In the EU it is further differentiated into E440(i) for non-amidated pectins and E440(ii) for amidated pectins (further details given below). The pectin world market has doubled the last 20 years and has a growth rate of approximately 3.5% per year. The annual sales volume of pectin is estimated at 34,000 metric tons (Brejnholt, 2007), and the global market value is estimated to be at least 400 million Euros (Savary, 2003).

Pectin is used for the production of jams and jellies, fruit beverages, confectionery products, bakery fillings, bread and dough production, dairy appli-

cations including acidified milk drinks and yoghurts, as well as pharmaceuticals and personal care products (Thakur, 1997; Rolin, 2002; Brejnholt, 2007).

Botanical origin

Pectin is found in the middle lamella and primary walls of plant cells. It is an important structural element giving both strength and flexibility to plant material. In Figure 32 the cellulose microfibrils are tied by cross-linking hemicellulose polymers. An important part of these are the xyloglucans that cross-link two or more cellulose microfibrils by hydrogen bonding. These cellulose-xyloglucan domains are embedded in an independent matrix of pectin polysaccharides.

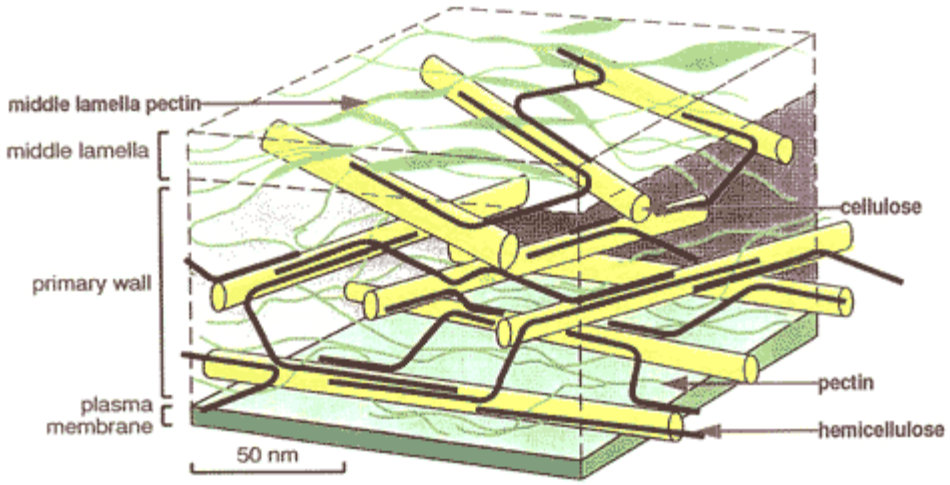


Figure 32 Simplified and schematic drawing of the arrangement of the three major structural polymers in a plant cell wall (reproduced from McCann, 1991).

Pectins present in the middle lamella are cross-linked to each other by Ca^{2+} and serve multiple functions. They limit the porosity of the cell walls, regulate the intracellular adhesion, add charged surfaces to the cell walls modulating the pH and ion balance and act as antigens signalling response towards symbiotic organisms, pathogens and insects (Carpita, 1993; Ridley, 2001; Willats, 2001).

Structure

Pectin consists of what is recognized as a family of oligosaccharides and polysaccharides which have common features. Pectin is one of the most

complex polymers known. It can be divided into three classes of polysaccharides all containing galacturonic acid (GalA) in various amounts. Pectin has been divided into "smooth regions" and "hairy regions". The smooth regions - called homogalacturonan (HGA) - consist of a linear chain of poly-α-(1→4)-D-GalA of 100-200 units where the GalA to a varying degree are methyl-esterified at the C-6 position (see Figure 33) and to a lesser degree acetylated at the O-2 or O-3 position (Schols, 2002). The hairy regions consist of rhamnogalacturonan I (RGI) and rhamnogalacturonan II (RGII). RGI consists of the repeating disaccharide [→4)-α-D-GalA-(1→2)-α-L-Rha-(1→] where Rha stands for Rhamnose. Attached to the rhamnose residues are one or more glycan side chains, primarily arabinan and galactan. The chain lengths may be up to 15 units and can be branched themselves. The chain lengths, branching and sugar composition of RGI can be highly heterogeneous.

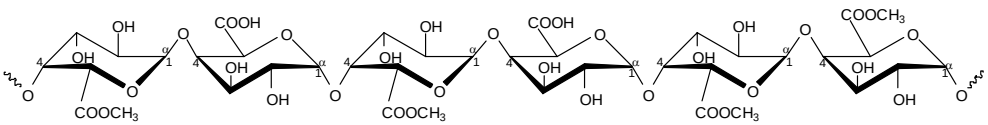


Figure 33 A fraction of a homogalacturonan (HGA) chain, which is partly methylated at C-6 positions.

The RGII has a backbone of HGA rather than rhamnogalacturonan contrary to what the name suggests. On the GalA residues, a complex arrangement of ordinary and very rare sugars is identified. RGII is nevertheless thought to have a highly conserved structure across different plants (Pérez, 2003). Until recently it was widely accepted to view pectin as having a "backbone" of HGA with hairy regions of RGI and RGII inserted as seen in part A in Figure 34. A newly proposed structure (B in Figure 34) has an RGI chain with HGA and RGII as long side chains (Vincken, 2003). Domains of xylogalacturonan, arabinan, arabinogalactan and apiogalacturonan may also be present. The exact composition and arrangement of HGA, RGI and RGII and other structural elements is very complex, and for this reason, no two pectin molecules are the same. The average molecular weight of pectin also varies a lot depending on plant origin and extraction conditions, but an average of 200 kDa as determined by liquid chromatography with light scattering detection is reported for commercial pectin (Rolin, 2002). This corresponds to a degree of polymerisation (DP) of ~1000 carbohydrate units. For a thorough discussion of the pectin structure, the reader is referred to (Thakur, 1997; Willats, 2001; Ridley, 2001; Schols, 2002; Willats, 2006) and the sources therein.

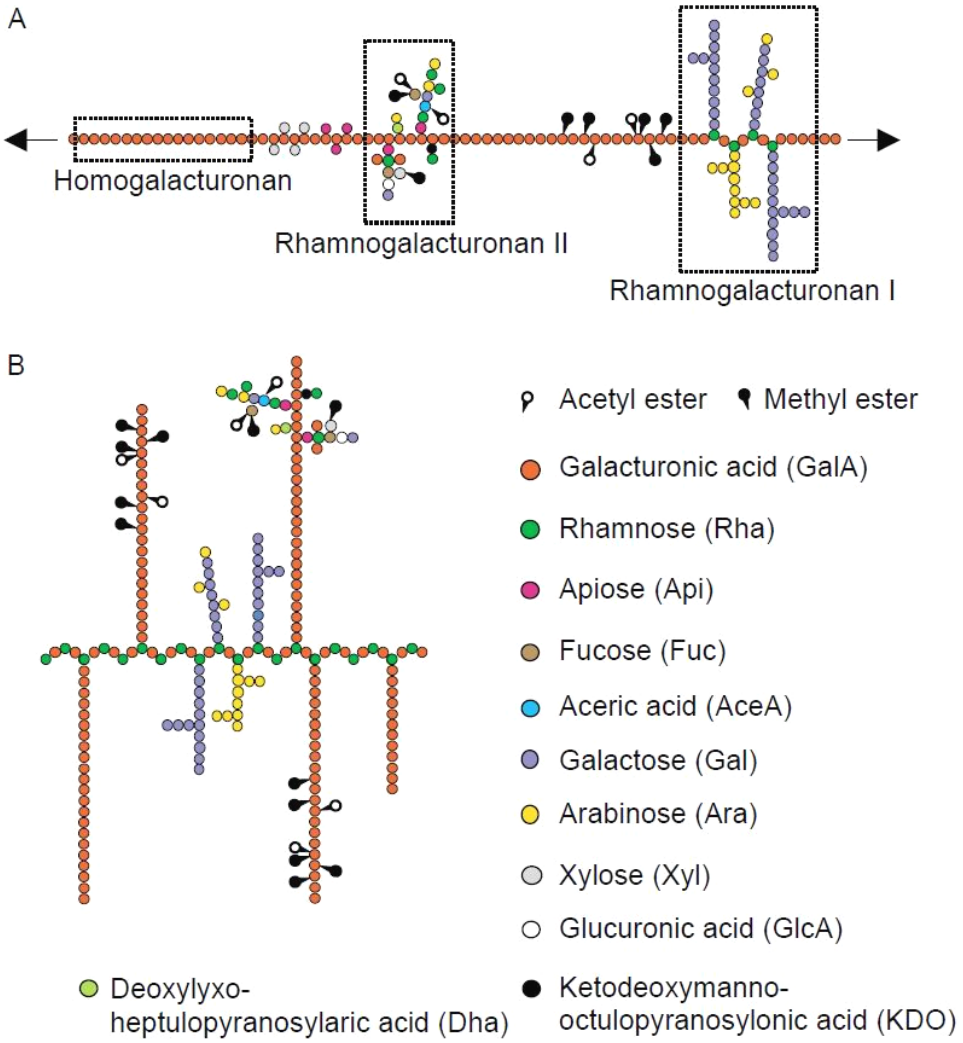


Figure 34 Two suggestions of the pectin structure: The structure in A consists of a "backbone" of HGA with hairy regions of RGI and RGII inserted. The newly proposed structure in B has an RGI chain with HGA and RGII as long side chains (reproduced from Willats, 2006).

The above described structure applies for *native* pectin also referred to as *protopectin*, which is pectin with the structure that is believed to exist in intact plants. In commercial preparations of pectin, part of the "hairy" regions is lost during industrial processing i.e. the side chains of neutral sugars and/or some of the RGI and RGII domains. *Commercial* pectin is therefore often simplified to partly methylesterified α(1→4) linked anhydrogalacturonic acid only. But this may be an oversimplification as the hairy regions

may be more or less retained. Recent research argues, that the HGA part of different acid extracted citrus pectins are quite homogeneous with respect to composition and molecular weight, and that the only significant difference were to be found in the RGI domains as only a minor RGII fraction was identified. In this research, a mass partition HG/RGI/RGII of ~88/11/1% was found (Yapo, 2007). In Table 3 the composition of typical commercially available pectins can be seen. There are still neutral sugars left and some phenolic and protein residues. Both types of pectins have been found to be slightly acetylated. The degree of acetylation (DAc) is estimated to 1.5 and 5.0% for lemon and apple pectins, respectively. The pectin extracted from apple contains more neutral sugars and more phenolics but less protein than lemon pectin.

Table 3 Composition of typical commercial pectins stated as percentage weight of dry matter (Kravtchenko, 1992). Values in parenthesis refer to degree of methyl esterification (DE) and degree of acetylation (DAc).

Component (%)	Lemon	Apple
Anhydrogalacturonic acid	76.4	60.8
Methyl ester groups (%DE)	4.4 (71.5)	3.6 (74.3)
Acetyl groups (%DAc)	0.26 (1.4)	0.72 (5.0)
Total neutral sugars	8.5	27
Proteins (N*6.25)	3	1.6
Total phenol	0.18	0.59
Ash	2.38	1.89
Total	95.1	95.9

Pectins are divided into subgroups depending on their degree of esterification (DE) which is the ratio of the esterified galacturonic acid groups to the total galacturonic acid groups. At times, the more specific degree of methyl esterification (DM) and the degree of acetyl esterification (DAc) are used, where DE ≈ DM + DAc. In commercially prepared pectin from lemon or apple the DAc is assumed to be negligible. When the DE is higher than 50% the pectin is called high methoxyl or high methylated pectin and abbreviated HM pectin. Pectin with a DE lower than 50% is referred to as LM pectin for low methoxyl or low methylated pectin.

Commercial pectin preparations typically have more than 70% by weight of galacturonic acid groups where, depending on the pectin origin and quality, 75% are methylated. The degree of esterification highly impacts the functionality of the pectin such as solubility, gelling ability and gelling tempera-

ture. HM and LM pectins combined are referred to as *pectinic acids* and the salts thereof are called *pectinates*. Polygalacturonic acid with no or almost no methyl ester groups (DE < 10) is called *pectic acids* and the salts thereof are called *pectates*.

If pectin is subjected to ammonia in aqueous/alcoholic suspension amidated pectin is produced. Some of the C-6 methyl ester groups are converted to amide groups, see Figure 35. The resulting degree of amidation (DA) is the ratio of the amidated galacturonic acid groups to the total galacturonic acid groups. Usually, HM pectin is amidated resulting in pectin with a DE < 50% and a DA > 10%. The resulting pectin is therefore called Low Methylated Amidated (LMA) pectin. This is in practice used for *all* types of amidated pectin, though pectins with a DE > 50% and a DA between 5 and 10% exist, and thus strictly should be referred to as HMA pectins. Pectins with DA > 25% are not considered food grade pectin and are not commercially available.

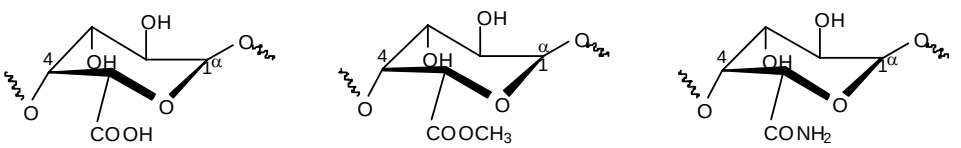


Figure 35 The three building blocks of the homogalacturonic backbone in a pectin molecule. Shown from left to right is a de-esterified, a methylated and an amidated galacturonic acid unit.

Physical/chemical properties

Pectins are soluble in water, but insoluble in aqueous solutions where gelling would occur if dissolved at a higher temperature and left to cool. The viscosity of the solution is related to the molecular weight of the pectin, DE, concentration of the pectin, pH and presence of cations. Monovalent cations will typically induce a decrease in viscosity whereas di- and trivalent cations increases viscosity. Viscosity, solubility and gelling are related properties. Factors that increase gel strength will increase the tendency to gel, lower solubility and increase viscosity. Pectin is a polycarboxylic acid due to the presence of GalA and is negatively charged at neutral pH. The dissociation of the individual carboxylic acid group is *not* independent of the other carboxylic acid groups. At 50% dissociation of the acid groups, the pH is likely to be between 3.5 and 4.5 (Speiser, 1945), so the pK_a value of pectin is there-

fore reported as between 3.5-4.5. The lower the pH, the less dissociated the carboxylic acids will be.

As seen in Figure 36, pectins in aqueous solution will undergo deesterification and depolymerisation. Best stability is found at pH around 4. At higher or lower pH both reaction rates will increase. In neutral and alkaline conditions the homogalacturonic backbone depolymerises by β-elimination. The ester group is furthermore prone to saponification (not shown in the figure).

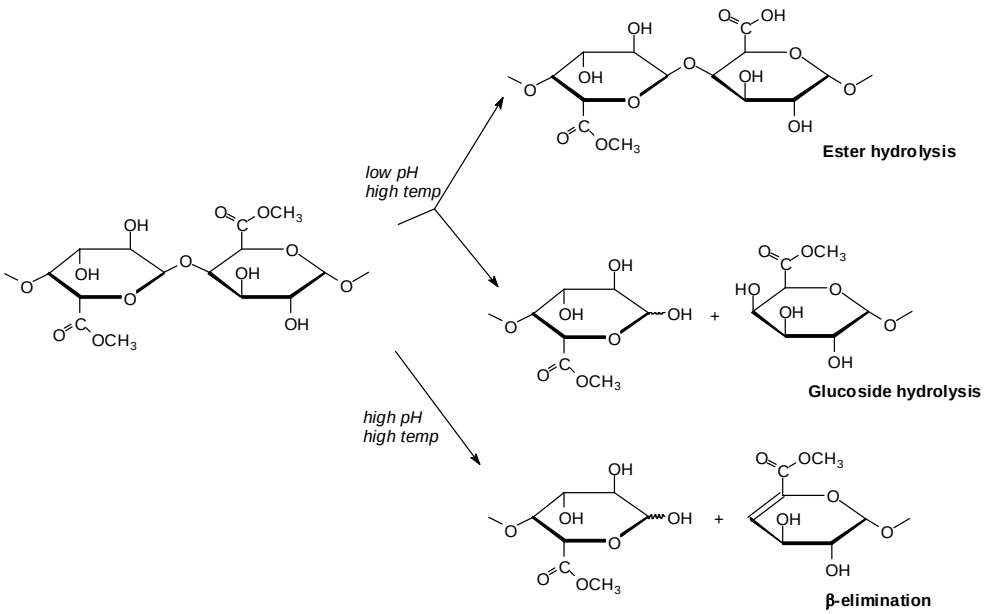


Figure 36 Common pectin degradation reactions.

Functionality

The primary functional property of pectin is its ability to form a gel in the presence of Ca^{2+} ions, sugar and acid. Pectin gels provide a desirable tender, short texture with excellent clarity combined with good flavour release properties (Engelsen, 1996). Several extrinsic factors affect the ability to form a gel: Temperature, concentration of pectin, pH, concentration of soluble solids (e.g. sugar) and the presence of ions, divalent ions such as Ca^{2+} in particular. The intrinsic factors of the pectin itself for gelling properties include molecular weight, the degree of esterification (DE) and amidation (DA), the presence of acetyl ester groups on O-2 and O-3 and possibly the presence of hairy regions. A gel may be regarded as a three dimensional network of cross-linked polymer molecules and as such an intermediate state between a solution and a solid. A pectin gel is an aqueous solution in which less than

4% pectin chains are cross-linked by hydrogen bonds, calcium bridges and other van der Waals forces and thereby trapping more than 96% water.
It is generally recognized that pectin forms at least two types of gel: Acid and calcium gels with different mechanisms of gel formation. Intermediates between the two also exist. The gel formation mechanism is depending primarily on the degree of esterification. HM pectin with a DE > 50% gels if the pH is below 3.8 and the concentration of soluble solids (typically sugar) is above 55% by weight. The function of the soluble solids is to lower the water activity, and thereby promote the pectin-to-pectin interactions. The low pH ensures that the carboxylic acids on the homogalacturonan backbone of the pectin are not dissociated thereby avoiding electrostatic repulsion between the pectin molecules. The result is that junction zones of hydrophobic interactions and hydrogen bonding between the methyl ester groups of different pectin molecules are stabilized. The non-covalent bonding of the junction zones forms the three-dimensional network of the gel. Increasing the degree of esterification and lowering the pH enhance the gelling ability (Speiser, 1946). Gels are normally made by mixing the hot ingredients above the so-called gelling temperature, the highest theoretical temperature that causes a gel to form in a particular system if kept infinitely long. When the system is cooled below a certain temperature gelling will occur after a certain delay which is short with pectin of high DE and longer with pectin of lower DE. For this reason HM pectin is divided into "ultra rapid set" (DE 74-77%), "rapid set" (DE 71-74%), "medium rapid set" (DE 66-69%), "slow set" (DE 58-65%) and "extra slow set" (DE 50-58%; Thibault, 2003). Table 4 gives an example of the required time before gelling has occurs in a typical gelling system. Pre-gelling is a state where an undesired, imperfect, lumpy gelling has occurred because the gelling process has been too fast. This leads to a broken set gel which exhibits syneresis (exudation of water after gel setting).

Table 4 Gelation of HM pectins with different degrees of methyl esterification at pH 3.0, soluble solids 65% and pectin concentration is 0.43%. Gelling time needed when cooled after boiling and held at indicated temperature (CP Kelco, 2003).

	Degree of methyl esterification (%)	**95°C**	**85°C**	**75°C**	**65°C**
Rapid set	**73.5**	60 min	10 min	*Pre-gelling*	*Pre-gelling*
Medium rapid set	**69.5**	*No gelling*	40 min	5 min	*Pre-gelling*
Slow set	**64.5**	*No gelling*	*No gelling*	*No gelling*	30 min

The calcium gelling of LM pectin, be it amidated or non-amidated, differs substantially from the acid gelling mechanism, as it is the carboxylic acids on the homogalacturonan backbone that is the functional group of the pectin. The pectin is still forming a three dimensional network, but the links involve ionic cross-links between two (dissociated) carboxylic acids and cations, calcium in particular. In the 1970's the "egg box" model was proposed (Grant, 1973). It involves two chains of homogalacturonan sufficiently deesterified so that it is possible for several Ca^{2+} ions to form coordination complexes with the carboxylic acids. Ca^{2+}-ions are then the "eggs" within two egg boxes formed by the homogalacturonan chains (Figure 37). As cations differ in valence number, ionic radii and electronegativity, their affinity for the electronegative gaps formed by the pectin differs. Ca^{2+} is the ion that has the largest influence on pectin performance apart from H^+ (Rolin, 2002).
LM pectins gel in the absence of sucrose. Therefore, LM pectins can be used to gel also low-calorie jams and jellies. Apart from DE and calcium concentration; molecular weight, presence of amide groups, pH, ionic strength and temperature also influence the gelling of LM pectins (Axelos, 1991).

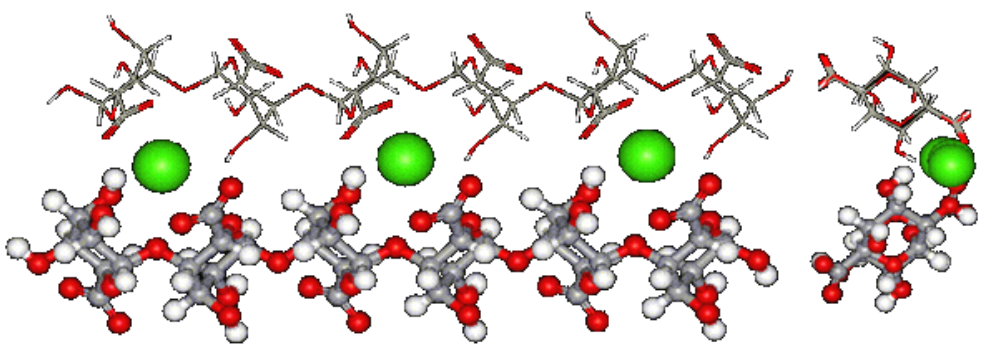


Figure 37 Representations (stick and van der Waals structures) of the chain-Ca^{2+}-chain associations of galacturonate chains; green circles represent calcium ions. On the right hand side the structure on the left is viewed down along the chains (reproduced from Braccini, 2001).

LM pectins (both amidated and non-amidated) are thermo reversible as opposed to gels made from HM pectins. Thermo reversible gels will re-melt upon heating and will be able to gel again upon cooling without a significant loss of gel strength. Table 5 sums up key functionalities of the various pectin types.

Table 5 Key features of high methylated (HM), non-amidated and amidated low methylated (LM) pectin (Brejnholt, 2007).

	HM	LM	Amidated LM
pH gelling range	2.5-3.8	2.5-6.0	2.5-6.0
Soluble solids range	55-85%	25-55%	5-65% (75%)
Gel setting temperature	25-90°C	40-100°C	30-70°C
Thermal reversibility	No	Yes	Yes
Gel texture at pH lower than 3.5	Firm	Spreadable with increase in firmness as pH is lowered	Semi firm, similar to HM but more rubbery
Gel texture at pH higher than 3.5	No gel is formed but viscosity is provided	Spreadable, slightly soft	Spreadable, similar to non-amidated LM

Recent studies reveal that the calcium gelling mechanism of LM pectins also influences the gelling of some HM pectins. Some HM pectins with a *blocky* rather than *random* DE distribution, i.e. having a group of neighbouring de-esterified galacturonic acids, have a higher gelling temperature in the presence of calcium. Such pectins are called calcium sensitive pectins, and have a different functionality (Rolin, 2002; Willats, 2006). It is a novel method for characterizing these pectins with a different *intramolecular* deesterification pattern that is the topic of **Papers II and III**. Gelling and gelling mechanisms are still highly debated in scientific literature. For a review of the discussion the reader is referred to (Thakur, 1997) or (Morris, 2007).
Pectins with high acetyl content, notably sugar beet pectins, do not form gels either with calcium ions or at high sucrose concentrations. They have been shown to be an effective stabiliser of oil emulsions (Williams, 2005).

Applications
Use of pectin in jams and jellies account for more than a third of the commercial production (see Figure 38). Before pectin was an industrial commodity, the natural pectin in fruits was used, and jams of low-pectin fruits were fortified by adding e.g. orange or apple to the recipe. Controlling the texture of the gel and setting temperature of jams and jellies are important to ensure uniform consistency, avoid fruit flotation and allow air bubbles to escape. Jellies and jams for baking will have to be thermally stable and are often required to withstand shear as many (industrial) applications require the jam to be pumped. Pectin is also used directly in bakery and other frozen products to retain water, minimise drip loss, improve bread volume and stabilise the food product e.g. the starch. Use of pectin for fruit toppings and fruit

preparations for yoghurt often calls for a soft LM pectin gel and care must be taken to avoid colour migration into the yoghurt's white mass phase. Confectioneries with pectin include low water content fruit jellies, foams and aerated confectionery products like marshmallows.

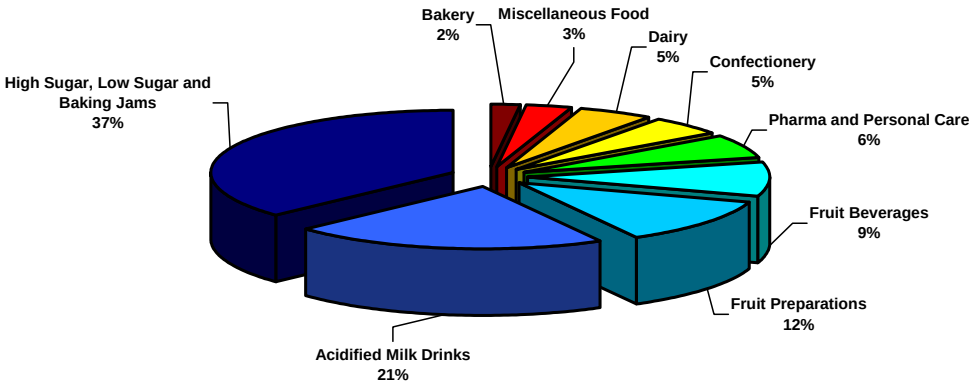


Figure 38 Estimated world market (by weight) of pectin applications (modified from Brejnholt, 2007).

Acidified and fermented milk drinks are the fastest growing application area of pectin. The negatively charged pectin molecules can bind to proteins carrying a positive charge and prevent them from coagulating when heated. The role of pectin is to stabilise the protein at low and intermediate pH and at the same time provide a desired mouthfeel. Careful control of the HM pectin is necessary to achieve this; therefore there is an increasing need for tailor-made pectin with tight specifications for the deesterification pattern (Willats, 2006). This necessitates the development of simple and rapid methods for quantifying this, as presented in **Papers II and III**. Other dairy applications utilize the natural calcium content in milk to create gels, spreadable or foamy milk or yoghurt desserts. In fruit beverages, pectin is used to increase viscosity and mouthfeel, mimic the presence of sugar or stabilize essential oils. The latter is an important use of sugar beet pectin. The non-food applications of pectin primarily include use in pharmaceutical and personal care products. Pectin is very skin friendly and highly biocompatible and is therefore used in cosmetics, lotions, breath strips, hair and skin care products as well as ostomy and wound care products. The functionality of pectin makes it suitable for pharmaceutical emulsions, anti-diarrhoea and anti-reflux products. Pectin is not degraded in the stomach and pectin is used for drug encapsulation for specific colon release formulations (Brejnholt, 2007)

Industrial pectin production

The first commercial production of a liquid pectin extract was established in 1908 in Germany, and the process spread rapidly to the United States, where a patent was obtained by Robert Douglas. This was followed by a rapid growth of the pectin industry in the United States and also, somewhat later, in Europe (IPPA, 2007). Douglas patents a "pectous concentrate…adapted particularly to be used in the making of jellies and which is adapted either for the use in a domestic manner by the housewife or in the more extensive manufacture of jellies, jams and preserved fruits or vegetables for public sale" (Douglas, 1913). Modern commercial pectins are almost exclusively produced as a precipitated dried powder for easy handling and improved shelf life. The raw materials are almost exclusively citrus peel (primarily lime, lemon and orange), and then some apple pomace and sugar beet pulp. As citrus fruit is the principal raw material for pectin production, it will serve to exemplify the process. The main producers of lime and lemon fruit are Mexico, India, Argentina, Iran, Brazil, and Spain in descending order, see Figure 39 for a complete overview.

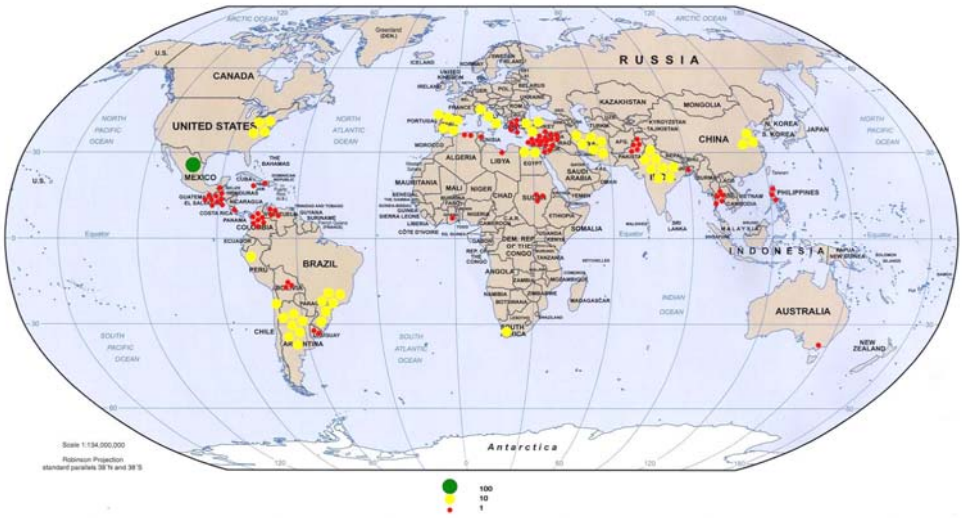


Figure 39 Global distribution of lemon and lime output in 2005 as a percentage of the top producer Mexico producing annually 2 million tonnes (FAO, 2007).

The raw materials are either used fresh or dried. As soon as the fruit is harvested, a decline of the fruit and pectin quality begins. At the juice processing facility, the fresh citrus fruits, in most cases, have their essential oils removed before juicing. The juicing leaves the albedo (see Figure 40), the fruit

interior (lamella, core and seeds) and the flavedo more or less intact. Following juicing, the quality deteriorates. The peel is chopped and washed to remove water and soluble solids and should be processed for pectin as soon as possible, or dried to extend the storage stability. An overview of the pectin production process is seen in Figure 41. The peel is suspended in hot (50-90°C) acidified water (pH typically 1-3) for 3-12 hours. Mineral acid (usually nitric acid) is chosen. The peel to water ratio is chosen, so that the resulting pectin concentration after extraction is 5-10g/l corresponding to ~ 0.5-1%. Depending on the extraction conditions and raw material, the pectin DE will vary between 80 and 55%. Large variation in the average molecular weight is also seen depending on the raw material and extraction conditions. The solution is filtrated, often in several steps to remove the solids. The resulting pectin thin juice may be ion-exchanged (to remove Ca^{2+} ions) and concentrated to pectin thick juice by evaporators or membrane filtration. The pectin is then precipitated in alcohol, typically ethanol or IPA (Isopropyl alcohol or 2-propanol). The pectin precipitate is washed in fresh alcohol, is pressed hard to a press cake, dried, milled, sieved and standardized to uniform gel strength usually by diluting with sugar which also facilitates the dissolution of the pectin for the user. Some types of pectin may contain suitable food grade buffer salts required for control of pH and desirable setting characteristics. An overview of some of the processing steps in pilot plant scale can be seen in Figure 42. If pectin with a lower DE than 55% is required, a further deesterification step is needed after the extraction. This step is done just before or right after the pectin precipitation. Amidation of LM pectins produces a firmer gel with less calcium than conventional LM pectins.

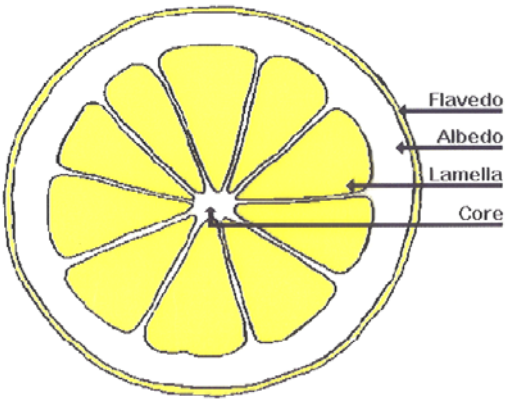


Figure 40 Different parts of lemon fruit.

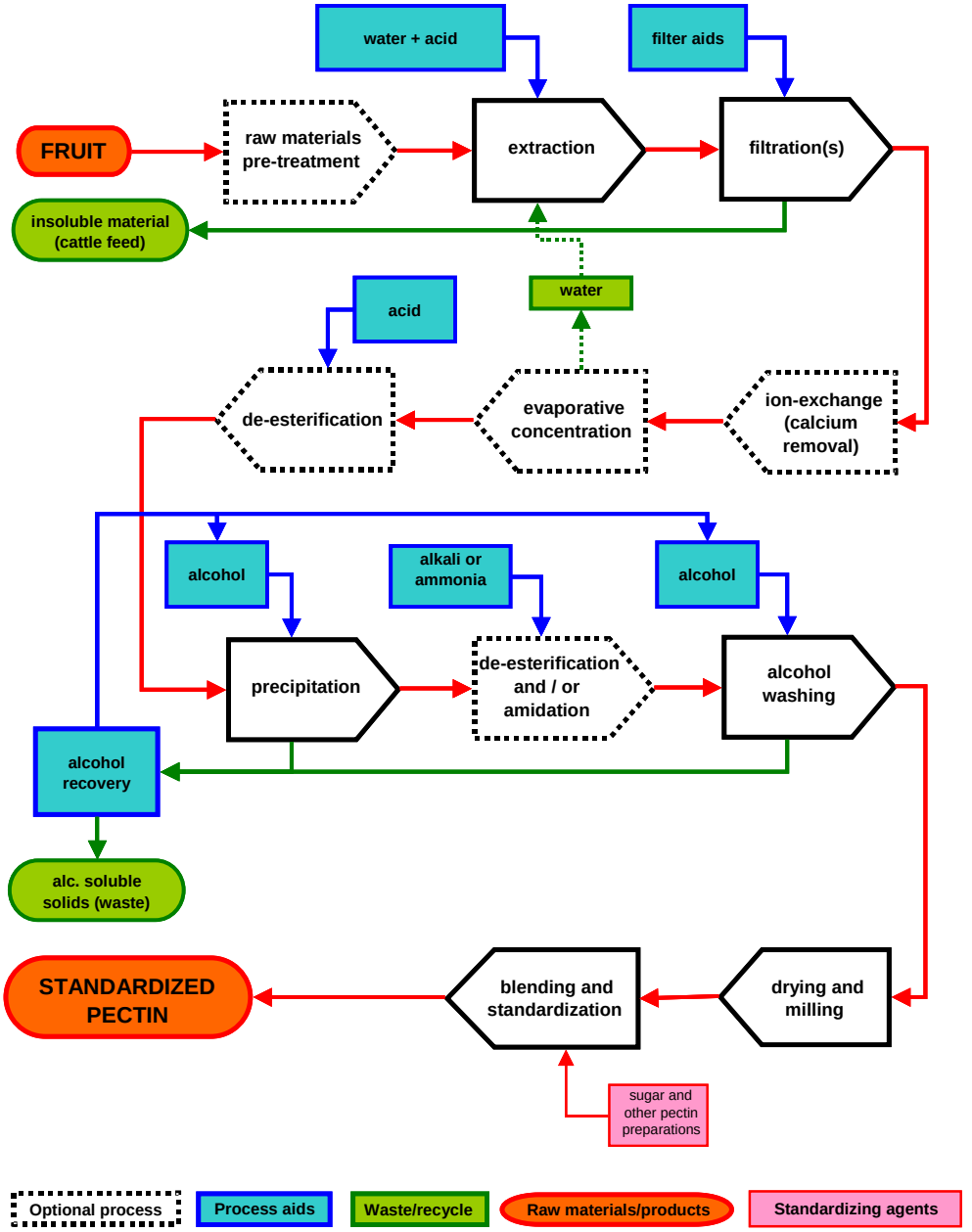


Figure 41 The pectin production process. If the deesterification or amidation steps are omitted, HM rather than LM or LMA pectin is produced (CP Kelco, 2003).

Figure 42 Various processing stages of pectin in pilot plant scale. From top left to bottom right: Raw material (oranges), chopped peel, extraction, filtration (filter aid), ion exchange (resin), precipitation, precipitated pectin and wet pectin press cake.

Establishing a PAT approach to pectin production

In order to establish process analytical technology (PAT) in a production environment, more than just the chemistry (e.g. spectroscopy) and the analytical side (e.g. the chemometrics) need to be understood and reviewed. In its broadest sense, a PAT approach comprises the full product chain from raw materials to consumer applications, as well as decision making layers involved in the process. This is illustrated in Figure 43. The product moves from right to left. Based on the raw material quality, certain raw materials are deliberately chosen for certain batches and appropriate process parameters are chosen. This will result in pectin with measurable fundamental qualities (e.g. DE, DA, molecular weight and intramolecular DE distribution). The pectin is however *not* sold or valuated based on the fundamental qualities but rather on its functional capabilities such as gel strength, calcium sensitivity and viscosities expressed in more or less elaborate grading or strength tests (Rolin, 2002). What matters for the customer is the pectin performance in their individual specific applications. For this quality, and not less important consistency, a pectin supplier is chosen, and value is given to the pectin by the customer (and thus pectin supplier).

For this reason it is a deliberate choice to have the raw material on the right hand side of the figure and double arrows between the boxes. Just as important as the product is moving from one side to the other in the chain, are information, knowledge, feedback and decisions moving in the opposite direction. It is far beyond the scope of this PhD thesis to elaborate on the entire pectin product chain in detail. Focus in this thesis will be on identifying some of the Critical Quality Attributes (CQA), the Critical Process Parameters (CPP) and identifying suitable techniques to measure specific process parameters and/or quality attributes. As already pointed out in the paragraph on functionality, the DE and DA and the intramolecular distribution of DE (and DA) are very relevant for the functional properties of pectin (Guillotin, 2006; Willats, 2006). It is an established fact that Raman, Infrared, and Near Infrared spectroscopy are appropriate techniques to characterize pectin and that the latter two are highly suitable to measure the overall DE and DA in pectin (Séné, 1994; Engelsen, 1996; Gnanasambandam, 1999; Synytsya, 2003). So far no papers have been published that clearly demonstrate that any of the mentioned techniques are able to determine the DE intramolecular distribution, however.

Figure 43 The pectin product chain from raw material to customer applications.

Figure 44 shows Fourier transform infrared (FT-IR) spectra recorded by the Attenuated Total Reflectance (ATR) technique of typical unstandardized powdered HM and LMA pectin samples. Tentative assignments of the fundamental vibration functional groups in pectin can be seen in the figure. Combination bands and overtones of the fundamental vibrations can be measured by NIR spectroscopy, and both FT-IR and FT-NIR have been established for more than a decade as off-line quality control (QC) methods to predict DE and DA in dry pectin powder at CP Kelco.

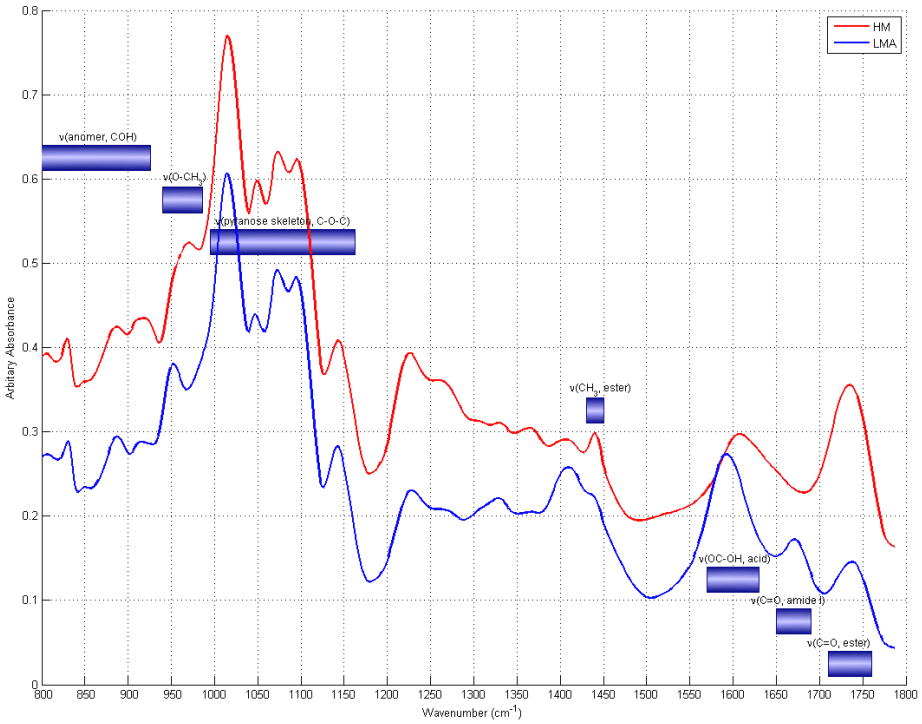


Figure 44 FT-IR-ATR spectra of typical HM and LMA pectin samples with tentative assignments partly based on (Engelsen, 1996) and (Synytsya, 2003).

The concentration of ammonia to be used in the amidation step is a critical process parameter for the control of DA. Likewise, the alcohol (IPA) concentration, the IPA/water ratio, is critical for optimal precipitation; therefore both IPA and ammonia are considered Critical Process Parameters, CPP's.

Furthermore, vast amounts of IPA is used throughout the factory in many processes and control of IPA concentration in process streams, and regeneration in distillation columns play a vital role for the overall energy efficiency of the process. Hence, monitoring and control of IPA streams have an impact on economy and environment. Ammonia *directly* relates to the DA of the pectin and its influence is well understood (Kim, 1978). IPA influences precipitation and thus *indirectly* influences secondary physical pectin quality attributes such as powder characteristics and specific gravity of the pectin. These relations are not well understood yet. **Paper I** describes the development of an on-line NIR based measurement point for the determination of ammonia (and IPA) with focus on real-time control of the ammonia for the amidation process.

Also developed during this project was an in-line and at-line system for the determination of DE and DA in wet HM and LMA pectin press cake by remote diffuse reflectance NIR. This system measures critical pectin quality attributes directly. The measurement is on the pectin itself rather than the amidation liquid which can be regarded as a process chemical. As described, the DE and DA of pectin can be controlled by manipulating the peel extraction conditions (i.e. pH, time, temperature) and e.g. ammonia concentration and temperature for the amidation reaction. The direct measurement of pectin quality attributes in-line and at-line makes rapid real-time feedback to the process possible. These results have not been published in scientific literature.

Papers II and III describe a novel method developed to measure the intramolecular DE distribution (i.e. the *blockiness*) of pectin. Actually, the method measures the Carboxylic Acid Distribution (CAD). Blockiness and intramolecular DE distribution are defined with respect to the ester groups, while this method measures the position of the unesterified galacturonic acids. In the presented setup it is an off-line method capable of analysing the end product. Contrary to other published methods, it involves little sample preparation and manipulation and can easily be taken to at-line. The CAD measurements can be used to assess the citrus peel used as raw material (by using the sample generated from a test extraction, which is done routinely to measure other raw material attributes). As presented in **Paper III**, the CAD analysis can be used for HM pectins (with a DE > 50%) only, but can in principle be extended to include LM and LMA pectins.

An overview of the above mentioned applications implemented on the LMA pectin process line can be seen in Figure 45.

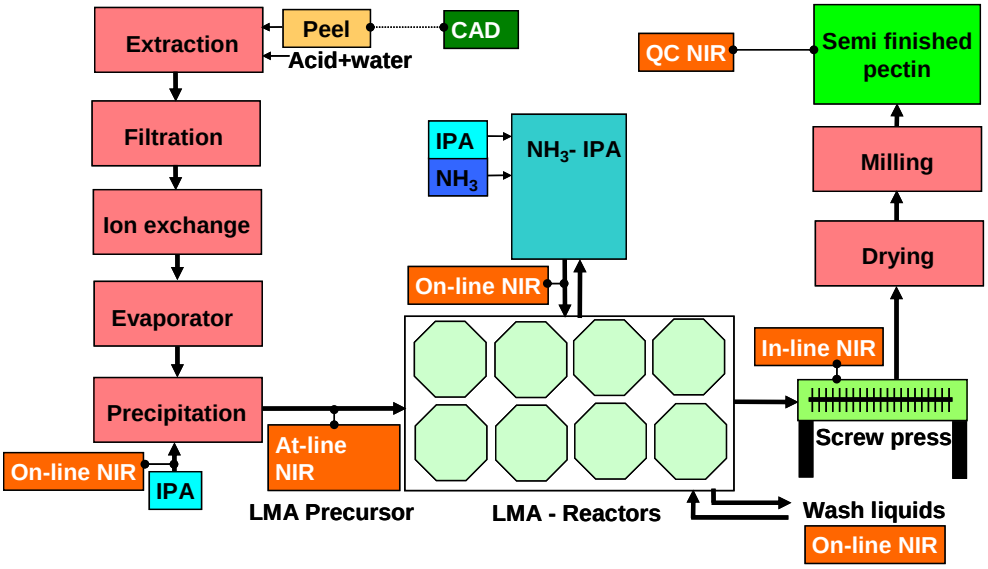


Figure 45 An overview of the NIR and CAD measurements as implemented on the LMA pectin process line.

Individually the applications may be seen as Process Analytical Chemistry applications, but together and combined with more traditional process measurements (pH, specific gravity of fluids, etc.) they provide the basis of a holistic overview of the pectin production. Then the individual measurements and applications unify to a technology which provides the control tools for the process and enables continuous improvement and knowledge management.

Controlling the amidation process

LMA pectin is typically produced by amidating HM pectin by ammonia in an alcoholic suspension (Bryant, 1949; Joseph, 1949; Kim, 1978; FAO, 2001). The amidation process at the CP Kelco facility at Lille Skensved, Denmark, takes place after precipitation on wet LMA Precursor, which actually is wet HM pectin press cake with a specified DE. The wet HM press cake is amidated in reactors in a suspension of ammonia, water and IPA. The process used to be controlled by hourly manual sampling and titration for ammonia in the main reactor tank. The dosing of ammonia was done manually, by turning a valve based on experience. No regular determination of IPA was done.

Fluctuations in the feed of ammonia to the amidation reactors, and loss of ammonia from the reactor during amidation, lead to less controlled amidation and thereby undesired product quality variation. It is difficult to closely monitor and act upon the fluctuations in ammonia concentration with sufficient accuracy. It could therefore be anticipated that an automated on-line, high frequency determination of the ammonia concentration during amidation, coupled with an automated dosing of ammonia, would result in a closer control of the process. This will in turn improve process capability, product consistency and improve adherence to product specifications. Feasibility studies were made in the laboratory. It was considered if it could be done and, if so, how it should be done. Also initial cost/benefits were contemplated. Several feasible paths were considered. Previously, an on-line UV-VIS based Flow Injection Analysis (FIA) system, based on the addition of indicator substances to the amidation liquid had been surveyed, but it did not perform adequately due to scaling of equipment, precipitation in tubing etc. It functioned in the laboratory but had insufficient on-line capability. NIR transmission spectroscopy was suggested as a possible more robust solution. A feasibility study revealed that it was possible to separate the main components of the amidation liquid by NIR spectroscopy (Figure 46).

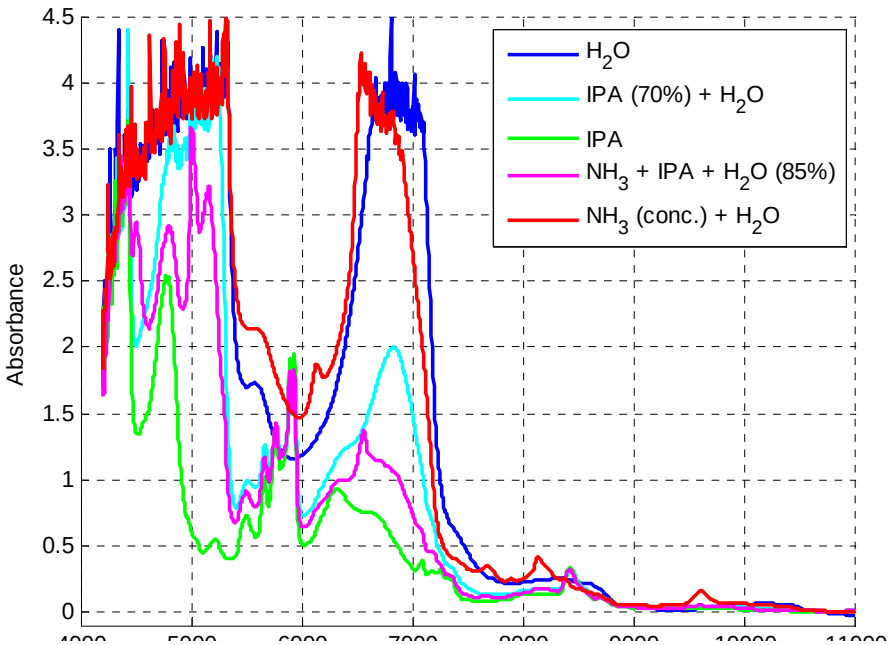


Figure 46 NIR transmission spectra of the main components in the amidation liquid and some rudimentary mixtures.

"Synthetic" (laboratory prepared) amidation liquid samples were made from a mixture design of IPA, ammonia and water. The advantage of NIR compared to IR is the higher energy of NIR light compared to IR light. This is a definite advantage in process transmission spectroscopy where the optical pathlength can be millimetres rather than micrometers. NIR light can also be transmitted through fibre optic cables at longer distances (30-100m), thus separating the NIR instrument from the process sampling area. A laboratory instrument was mounted with a fibre launcher, so measurements could be done with fibre optic cables in order to mimic the process spectroscopy as closely as possible. Later, the laboratory environment was substituted with the production environment in an experimental on-line setup, where additional design features were tested in-situ and invaluable experience was acquired regarding the process sampling environment. An option was put forth to use a new side stream sample cell with Teflon® tubing instead of the traditional stainless steel flow cell. The design is simple: A Teflon tube is mounted in a suitable housing. Teflon does (almost) not absorb near infrared light and the optical path length is adjustable by spacers. Teflon is chemically robust and only the tube is in contact with the product in adherence with a no glass policy in food production environments. The working range is up to 6.9 bar and 100°C, and the measuring cell can be mounted in a waterproof house.
The sample acquisition and chemometric modelling are thoroughly documented in **Paper I** and the preceding chapter on chemometrics in this thesis. Careful choice of spectral pre-processing and wavelength selection could eliminate a temperature dependency, which had to be solved or the prediction performance would be impaired. Similar successes are reported elsewhere in scientific literature on a somewhat similar data set (Wülfert, 2000). When the on-line NIR ammonia predictions were compared to the manual titrations done by process operators, the results were surprising. The process operators routinely analyzed a sample approximately once an hour to monitor the ammonia concentration. If the level was assessed as low or if a new amidation was forthcoming, the operators would feed ammonia to the main tank by manual operation, based on process experience. The operational effect would be evaluated from the next titration one hour later, as titrating a sample immediately after dosing would give an "unreliable" (i.e. out of specifications) result. Figure 47 shows the results of the NIR predicted concentrations superimposed on the process titrations. When only the process titrations are considered, the process appears stable and within the specified ammonia concentration set point. The NIR predicted results, however, re-

veal that the dosing of ammonia by the operators results in strong concentration increase (reaching levels of more than 15% above target point), which is rapidly consumed in the amidation reaction. The magnitude of the concentration peaks and the rapid consumption were unknown to process engineers before implementing the process analytical measurement system. For reasons of confidentiality, the concentrations of IPA and ammonia have been rescaled to 0-100% following the convention of **Paper I**. This has the inherent advantage that Root Mean Squared Errors will be expressed in dimensionless % of the calibration range, which enables calibration performance comparison to other applications.

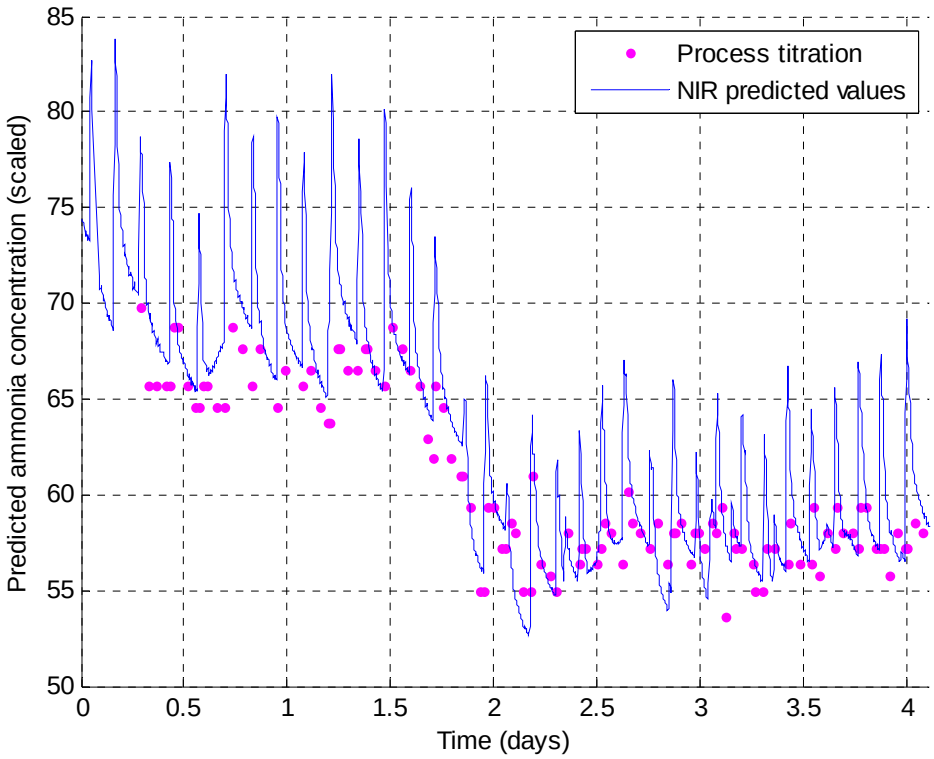


Figure 47 NIR predicted values superimposed on manual ammonia titrations before and after a process set point change taking place at day 2. The hourly titrations do not capture the full dynamics of the manual ammonia dosing process.

The NIR predictions are extremely precise and very stable from one measurement to the next (best observed in a time span where no dosing of ammonia occurs). This implies that the spectroscopy plus process interface as well as the PLS model are very robust, and noise levels are very low.

After an evaluation period, the NIR predictions proved reliable and process control engineers phased in an automated dosing system based on the NIR predictions. An automatic valve, capable of dosing ammonia in short bursts, replaced the manual valve. The automatic dosing was introduced slowly over a period of 14 days, in order to gather experience and not to upset production quality. Figure 48 shows NIR predicted ammonia values from the production during that period. It is clearly observed that the ammonia concentration is much more in control, being kept at an almost fixed level with near continuous automatic dosing, rather than manual dosing every other hour. Concentration set points can now be readily changed within minutes, compared to several hours before the introduction of the new system.

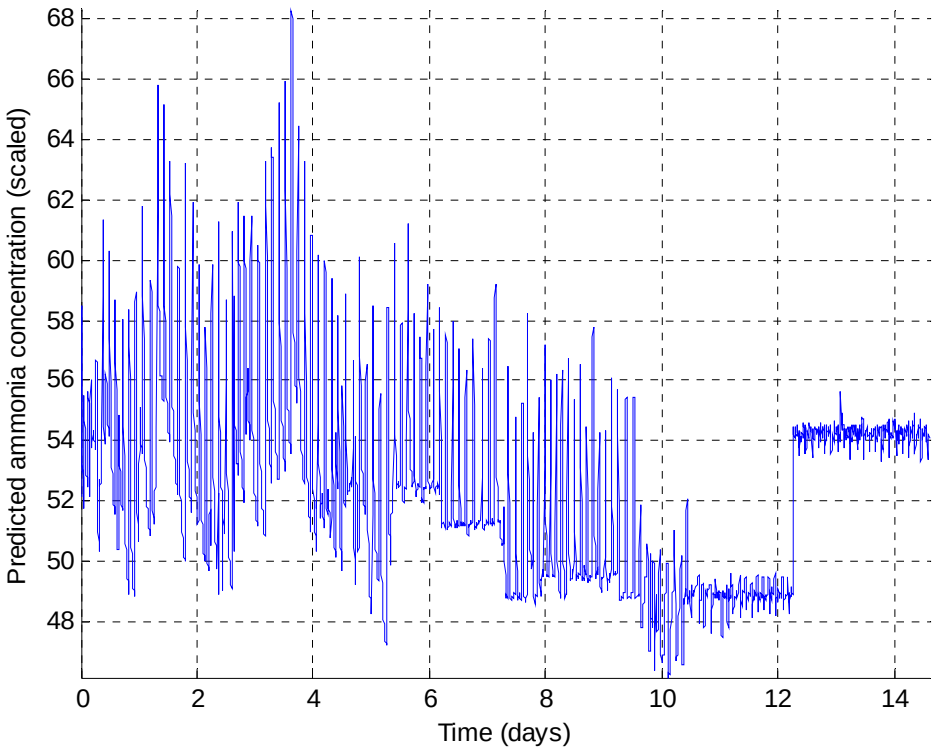


Figure 48 NIR predicted ammonia concentrations during a period of 14 days where on-line control and dosing of the ammonia concentration was implemented. A process set point change was imposed after 12 days and the dosing system responds immediately.

The benefits of the improved control of the ammonia dosing system have been numerous. It is now possible to regulate the concentration of ammonia

very precisely and reproducibly. This removes an important source of variation from the amidation reaction, and at the same time it offers a possibility to make minute changes to ammonia concentration to improve product quality. If troubleshooting of the production becomes necessary, process engineers can assume or quickly verify that the ammonia concentration has been constant.

While **Paper I** describes the measurement, determination and control of a critical process parameter, it does not measure the pectin itself. Nor is it possible to control or predict the DA of the finished LMA pectin solely by adjusting the ammonia concentration. Other factors such as amidation time, temperature, initial DE of the HM pectin precursor before amidation and physical characteristics of the wet press cake before and during amidation may influence the process. For this reason, an in-line and at-line system for the determination of DE and DA in wet HM and LMA pectin press cake by remote diffuse reflectance NIR has been developed. Owing to the fact that off-line NIR determination of DE and DA in the QC laboratory on finished, dried and milled pectin powder has been implemented for more than a decade, a NIR solution was chosen. A fibre optical probe was designed and fitted in a screw press in the production. From Figure 49 it can be gathered that the process environment is quite harsh, and it literally calls for implementations of robust methods in all senses of the word!

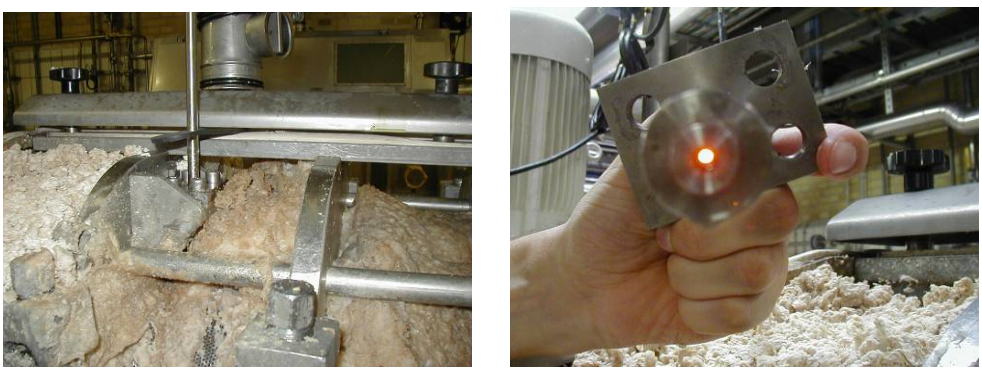


Figure 49 Fibre optical probe embedded in a pectin screw press (left). The lid is off and the brownish mass is wet pectin press cake. On the right is the probe with the fibre bundle in the middle on the tip. Visible light accompanies the NIR light from the light source.

These results have not been published in scientific literature, but pertinent results will be discussed in the following chapter.

Measuring the Carboxylic Acid Distribution

The exact distribution of ester groups on the pectin backbone is a property that can only be defined for an individual pectin polymer, and this information is of no commercial interest as it is not possible to extract or produce identical pectin polymer molecules from natural sources. On the other hand, a method for measuring average intramolecular distribution of ester groups in bulk pectin is highly desirable. In the quest for new and improved pectin functionality, the design, control and manufacturing of tailor-made pectins are important (Willats, 2006).

Two different types of pectin esterases are known to deesterify pectin in either a predominantly "Random" or "Blocky" fashion. The terms Random and Blocky are loosely defined terms, but it has been suggested to define a blocky group as a sequence of four or more deesterified galacturonic acids next to each other in the pectin carbohydrate backbone (Daas, 2000; Limberg, 2000), while Random groups in this context may cover truly random or more systematic (sequential) deesterification patterns, see Figure 50.

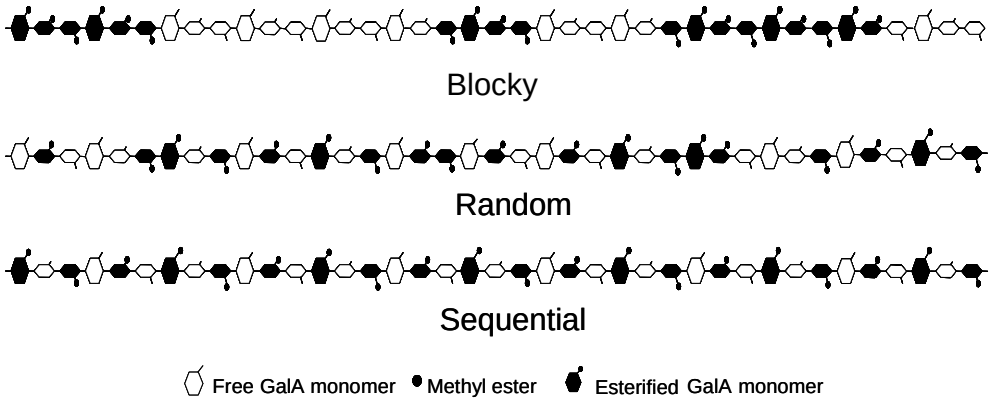


Figure 50 Idealized diagram of three sections of pectin backbone (GalA = galacturonic acid). The pectins are deesterified 50% but the distribution of ester groups is different (modified from Winning, 2007).

Pectin esterase (sometimes referred to as pectin pectylhydrolase EC 3.1.1.11, pectin methylesterase, pectin demethoxylase, pectin methoxylase or pectase) de-esterifies the methyl esters of the carboxyl groups, producing methanol and low methylated pectin. The enzyme is produced by higher plants and micro organisms. Pectin esterases derived from plants (Solms, 1955; Miller, 1971) are thought to attack either at the non-reducing end or next to a free carboxyl group and to proceed along the molecule by a single chain mecha-

nism, creating blocks of unesterified galacturonic acids that are very calcium sensitive (Kohn, 1968). Irregularities in the pectin galacturonan chain inhibit the activity of the pectin esterase. Irregularities could be acetylated monomers, the occurrence of carbohydrate side-chains (e.g. hairy regions) or ester groups that are transformed into amides (Solms, 1955). Pectin esterase is highly specific for the methyl ester of polygalacturonic acid. The rate of pectin deesterification depends on chain length; trimethyl trigalacturonate is not attacked at all (McCready, 1954). Most fungal pectin esterases differ from plant pectin esterases by obeying a multi chain mechanism, removing methyl groups at random (Ishii, 1979).

Most methods for determining the distribution of ester groups or the inverse property, the free carboxyl group distribution, involves enzymatic digestion of pectin. Tuerena et al. suggests blocking the free carboxyl groups by glycolation and do a hydrolysis of the methylesterified regions with a mixture of pectic enzymes. The hydrolysis products are separated from the glycolated regions on an ion exchange column and the resulting oligomers are analyzed (Tuerena, 1982). De Vries et al. degrade pectin with purified pectin lyase and separate the fragments by gel permeation chromatography and HPLC (de Vries, 1986). The pioneering work was followed up by Daas et al. using high-performance anion-exchange chromatography (HPAEC) introducing the term 'degree of blockiness' (DB) which is the total amount of non-esterified GalA liberated as the percentage of the total number of non-esterified GalA present in pectin (Daas, 1999). This approach is still used extensively (Guillotin, 2005) but can be very time consuming because of the deesterification and digestion steps involved. As an alternative Capillary Electrophoresis (CE) have been explored as a tool (Jiang, 2001; Guillotin, 2007) and also ^{1}H NMR and ^{13}C NMR has been used to quantify the distribution pattern (Grasdalen, 1988; Westerlund, 1991; Rosenbohm, 2003; Winning, 2007). Needs et al. degrades pectin chemically, rather than enzymatically, in six steps before analyzing the using high-performance size exclusion chromatography (HPSEC; Needs, 2001). Recently sophisticated molecular methods have found their way to the analysis of food matrices such as matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) and Ion-Trap mass spectroscopy (MS) for characterization of the pectin digests. Use of monoclonal anti-pectin antibodies with specificities for pectin side chain and homogalacturonan backbone domains have increased and combined with high throughput microarray technologies (Willats, 2006). Common for most of the above mentioned methods are that most involves many preparative steps and can be very time consuming.

How "Blocky" or "Random" pectin is, is not easily expressed as a single number, the before mentioned degree of blockiness is the most successful. A method will also have to take into account that the average total degree of esterification (DE) will vary from sample to sample. As several "highly multivariate" questions are posed, a multi-way, multivariate method is suggested to utilize the established multi-way advantage. But in order to calibrate the method properly, reference pectins with known deesterification patterns have been made. As natural citrus pectins have a DE typically lower than 80% when extracted, almost 20% of the galacturonic acid groups are already deesterified in an unknown pattern. Therefore it is necessary to remethylate the pectin to a DE as close to 100% as possible. A DE of 93.8% was achieved in the laboratory. The remaining 6% were impossible to remethylate and it is possible that the remaining 6% are neutral sugars in between the galacturonic acids, so that the homogalacturonan domains are flawed or remains of RGI (see Figure 34).

The remethylated pectin is deesterified in a controlled fashion with plant and fungal esterases inducing blockwise or random deesterification patterns, respectively, see the upper half of Figure 51. If in theory the pectins could be deesterified completely (to a DE of 0%), the resulting polymer would be a pectate and similar regardless of the fashion in which it was deesterified. In reality the enzyme activity drops to 0 before the pectin is completely deesterified. Somewhere in between a DE of 100% and 0% the two types of pectins are most different with respect to functionality. It does not have to be at a DE of 50% as suggested in Figure 51.

An exploratory experimental study was set up using a pectin binding dye, in which – during a controlled dilution gradient in a liquid system – a diode array detector (DAD) has recorded UV-VIS spectra, yielding a three-way matrix of samples × wavelength × dilution time. It is readily observed that transmission spectra of dye and pectin in different concentration ratios enhance the information, as the 2D spectral fingerprints (landscapes) of the pectins are very different, see the bottom half of Figure 51. The feasibility studies were also presented (**Poster I**) at the 9th Conference on Chemometrics in Analytical Chemistry, Lisbon, 2004.

In principle, the ratio of blockwise to randomly distributed acid groups could be identified in a known amount of pectin without having a dilution gradient, but such a method would depend on the DE. The DE could be measured by conventional methods e.g. NIR or titration. But in order to quantify truly unknown samples, multi-way data will have to be collected at different concentrations.

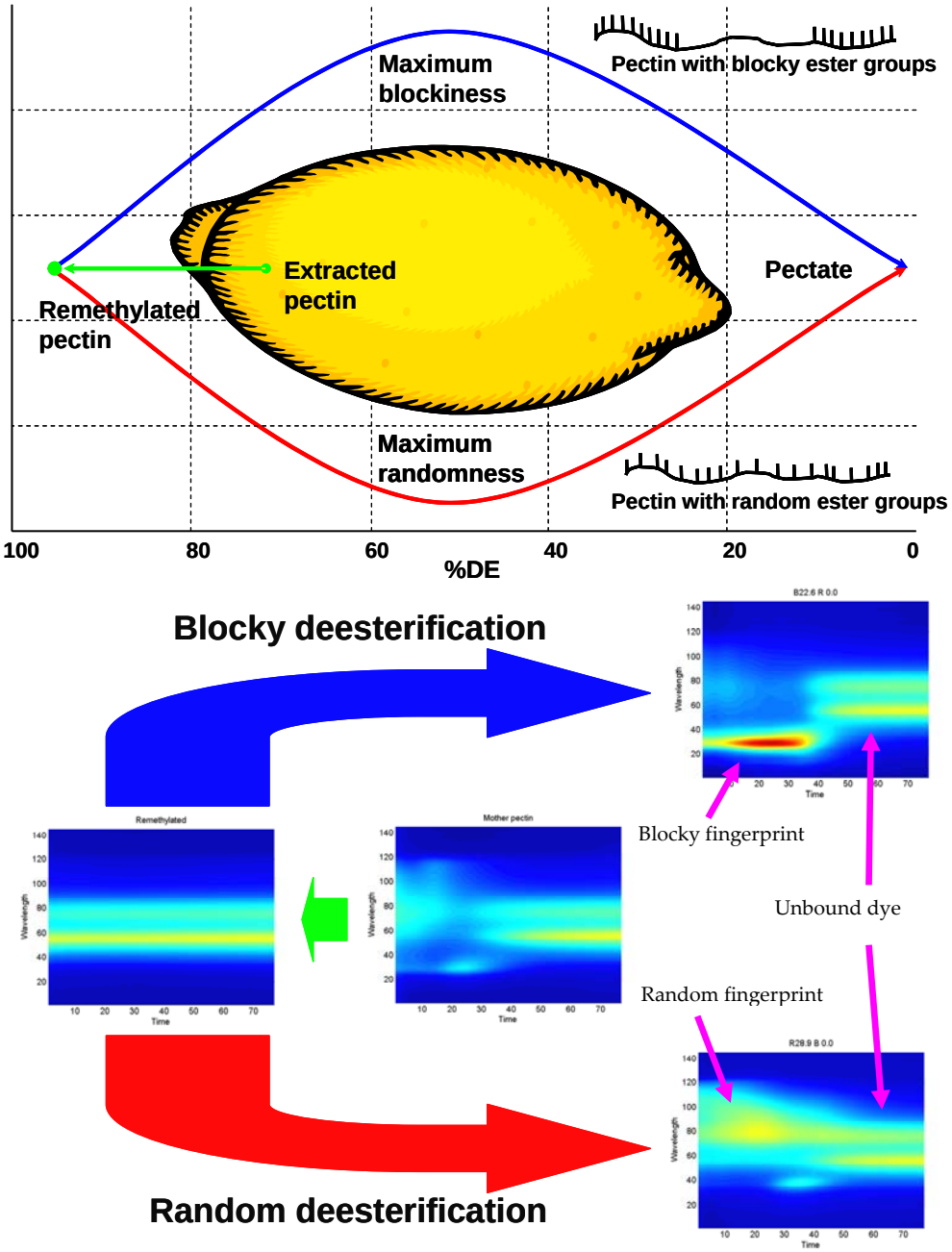


Figure 51 The sample set is generated from a "Mother pectin" extracted with a DE of ~72%. The pectin is remethylated to DE ~94% and subsequently deesterified controlled with pectin esterases known to deesterify in either a random or blockwise fashion. The wavelength × dilution time landscapes (seen at the bottom half of the figure) of two resulting samples (with a DE of 71 and 65%) have distinct features (pointed out) reflecting the different deesterification patterns.

A 0.5% (weight by volume) solution of pectin is injected by a sample loop connected to an injection valve into a carrier stream flowing continuously and containing a dye in a fixed concentration. The stream leads into a continuously stirred tank reactor with a fixed volume, large compared to the flow rate of the carrier stream. The reactor is initially filled with the carrier stream containing the dye. The injection volume of the pectin solution is small compared to the reactor volume, so the concentration of dye can be assumed to be constant. The exit from the reactor leads to a flow cell where the UV-VIS transmission of the stream is measured. After the measurement, the stream is led to waste. The reactor gives the system a large dead volume that retains the injected pectin, which is only slowly washed away by the carrier stream, see Figure 52.

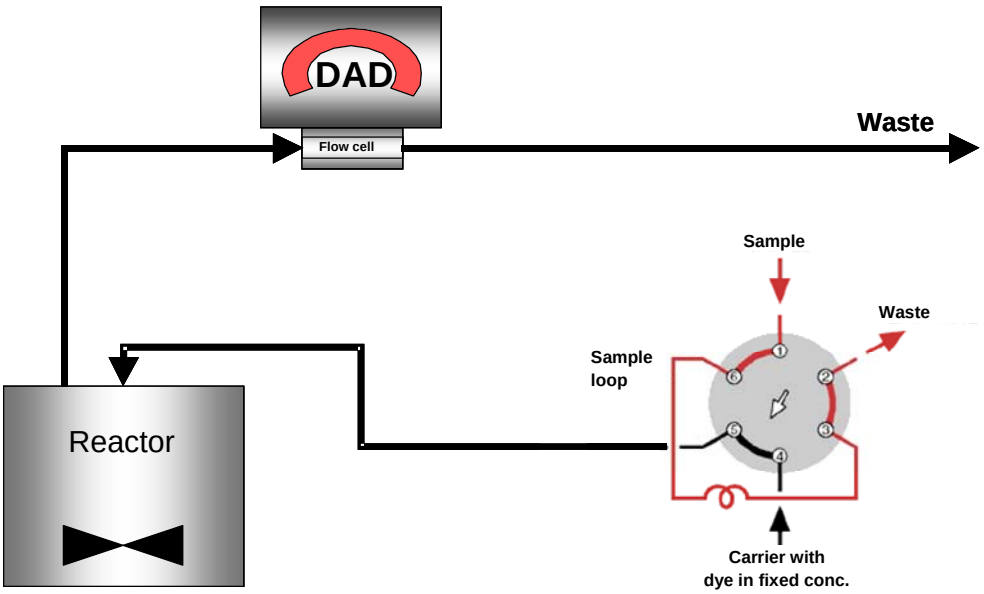


Figure 52 Overview of the setup to measure the Carboxylic Acid Distribution (CAD). On the bottom right hand side of the figure is a 6 port rotating valve that sends either the carrier or the contents of the sample loop to the reactor.

The dye is known to bind to the pectin irreversibly (within the time frame of the experiment) with a high affinity, and changes conformation (i.e. absorbance) depending on whether it is bound or not bound to the pectin. It is observed from initial studies that the dye, when bound to the pectin, also changes conformation depending on whether the neighbouring galacturonic acid on the pectin backbone contains an ester group or an acid group (with

another dye molecule attached), i.e. the dye conformation is different if the molecule binds to a blocky or to a random pectin environment.
The concentration of pectin injected is initially high compared to the concentration of dye in the reactor. It is therefore assumed that the dye binds preferably to the easily accessible sites on the pectin. As the pectin (with dye attached) is washed away from the reactor by the continuously flowing carrier stream (still containing the dye in a fixed concentration), the concentration of pectin decreases exponentially. As the concentration of pectin drops, the dye will also bind to the less accessible sites on the pectin. At the end of the sample run, the pectin concentration will be close to zero and negligible. Hence, only the unbound dye will be spectroscopically active at that stage. The different spectral landscapes that are recorded can be reduced to three basic features. First the spectra fingerprint of the random acid groups, then the blocky acid groups and at last the unbound dye appear. **Papers II and III** describe the sample set, the sample naming conventions and the advanced multi-way chemometrics in detail, but the conclusion is that the spectral profiles can be modelled, quantified and related to the pectin de-esterification pattern as suggested in Figure 53.

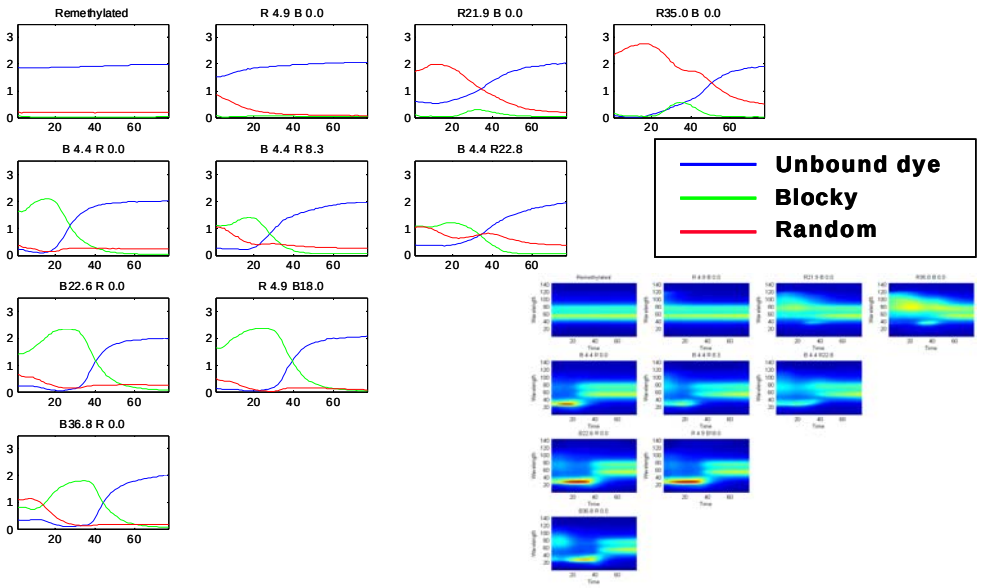


Figure 53 Resolved concentration profiles for the unbound dye, the blocky and the random spectra profiles. The concentration profiles have been resolved by a three component MCR-ALS model with non-negativity constraints. The actual recorded spectral landscapes can be seen inserted for reference.

Benefits from implementing PAT (in pectin production)

When implemented, PAT is expected to reduce the risk of scrap and recalls, reduce production cycle times and enhance capacity utilization, thereby in the long run, deducing product development time (Hussain, 2002). This is due to the emphasis not only on process analytical chemistry tools, but also information management tools, feedback process control strategies, product and process design and optimization strategies (Workman, 2005).

Integrating information from several sensors in- or on-line

As an example of the benefits from implementing a unified PAT strategy can be that a synergy can arise between different sensors placed in- or on-line. In Figure 54, the on-line predicted and controlled ammonia concentration in the amidation liquid and the in-line predicted %DA values of the wet pectin press cake *after* the amidation, are seen. See Figure 45 for sensor placements.

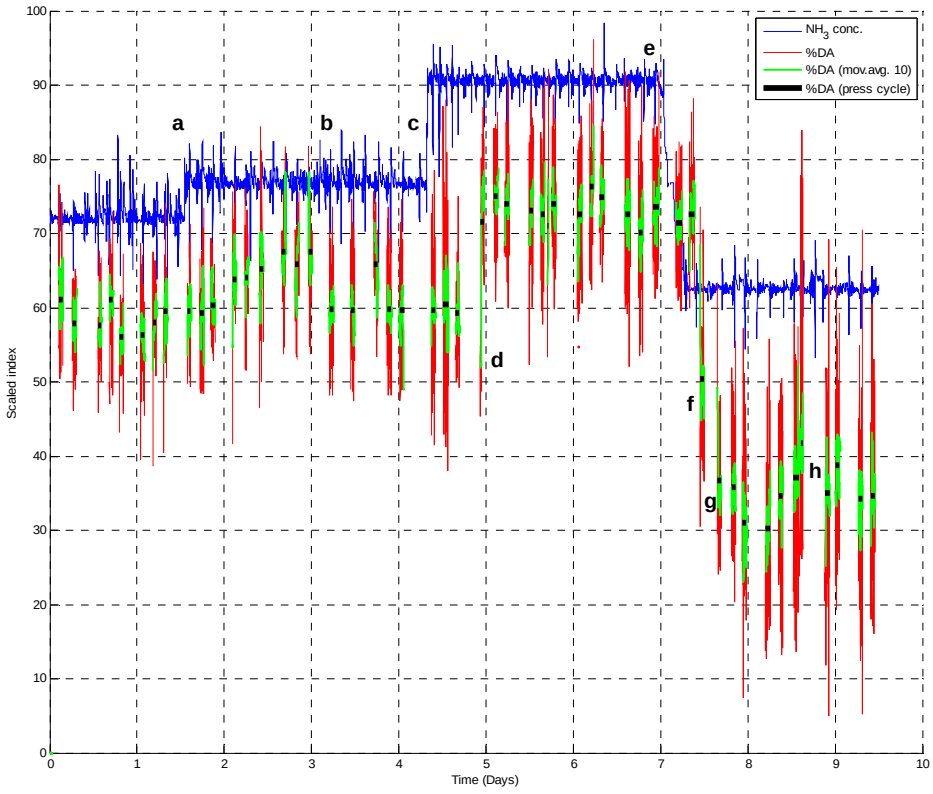


Figure 54 On-line ammonia prediction in the amidation reactors (blue) and in-line predicted %DA in pectin press cake (red, and moving average of 10 %DA observations is green, average of entire press cycle is black) in the screw press following the amidation. Both have scaled values on the y-axis. See the text for further details.

In-line %DA data are only collected when the screw press is running i.e. when the amidation reactors are emptied. The measurement to measurement variation on the %DA determination is quite large, but the moving average of the on-line measurements correlates well to the reference measurements. This is an example of the in-line or on-line advantage in the case that the measurement is "imprecise but accurate". Since the in- or on-line measurements are fully automated, a high number of measurements can be taken which improves the accuracy if the measurements are not biased relative to the imprecision. Also immediately noticed is that lagging is evident. The pectin press cake is first subjected to ammonia in the amidation reactors, then several washing steps follows before the pectin is pressed in the screw press. The highest cross correlation between the ammonia concentration and the %DA is 0.80 at 9 hours and 50 minutes of lag. This means that material from the amidation reactors (several reactors running batch-wise) can be seen on average about 10 hours later in the screw press as the batch is emptied. This corresponds to approximately 3 screw press cycles. At time point "a", as indicated on Figure 54, the ammonia set point is increased in order to ensure better adherence to pectin functionality specifications. The %DA is seen to increase 3 press cycles later reflecting the lag. At time point "b", the %DA falls back slightly. This is due to a change of the raw material (another citrus peel lot has been taken in use) at an earlier time in the extraction process that affects the resulting %DA, even though the amidation conditions are unchanged. At time point "c" the campaign is finished, and a new pectin recipe with higher %DA is to be produced and the ammonia concentration is increased significantly. Some residue of the previous campaign can be seen at time point "d" at the beginning of a screw press cycle just before the %DA increases. This is carry-over material from the previous campaign. At time point "e" the ammonia concentration is lowered as yet another pectin recipe is to be produced. The product produced at set point "f" is an intermediate product between the second and third pectin recipe. It resembles %DA-wise, the first pectin recipe produced rather than the second or third pectin recipe. At time point "g" some more carry-over material can be identified. At time point "h" some batches with an apparent larger measurement imprecision can be identified. This could for example be because the pectin in reality is more inhomogeneous or because process or sampling conditions in-line in the screw press are varying more during the particular screw press cycles. In conclusion a lot of information can be gathered when combining the sensor responses rather than keeping information from different unit operations

separate. This gives a holistic view into the production on-line, and visualises ways to analyze, improve and validate the process.

Detecting abnormal process conditions

As mentioned in the chemometrics chapter, one of the benefits of using multivariate statistics is the extensive diagnostic and error measures that are available. One of these error measures is the "F-ratio". The F-ratio relates the spectral residual of a sample to a model reconstructed spectrum i.e. a reconstructed spectrum using the loadings of the PLS model used for the calibration, and the computed score values of the sample spectra in question. The spectral residual is the difference between this spectrum and the actual prediction spectrum and is calculated as:

$$E_{reconstructed} = \sum_{j=1}^{J} (A_{original,j} - A_{predicted,j})^2 \quad (17)$$

where J is the number of wavelengths in the spectrum, A_{original} are the original spectrum absorbances, and $A_{\text{predicted}}$ are the model predicted spectrum absorbances. For spectral residuals, the F-ratio is calculated as:

$$F_{ratio,new} = \frac{(K-1)E_{reconstructed}}{\left(\sum_{k=1}^{K} E_{calibration\ set,k}\right)} \quad (18)$$

where K is the number of samples in the calibration set, $E_{calibration\ set,j}$ are the spectral residual value of sample k in the calibration set and $E_{reconstructed}$ are the spectral residual value of the sample in question (Haaland, 1988).

The F-ratio should normally be low (typically less than 10, depending on the application). An increase in the F-ratio may indicate abnormal measurements, and the cause may be the product itself, the sampling system or the process analyzer. In Figure 55 an increase of the F-ratio during 10 days of both the ammonia and IPA can be seen. At the same time the precision of the predictions is deteriorating. The measurements are from of the auxiliary tanks mentioned in **Paper I**, not the main tank responsible for the on-line dosing of ammonia.

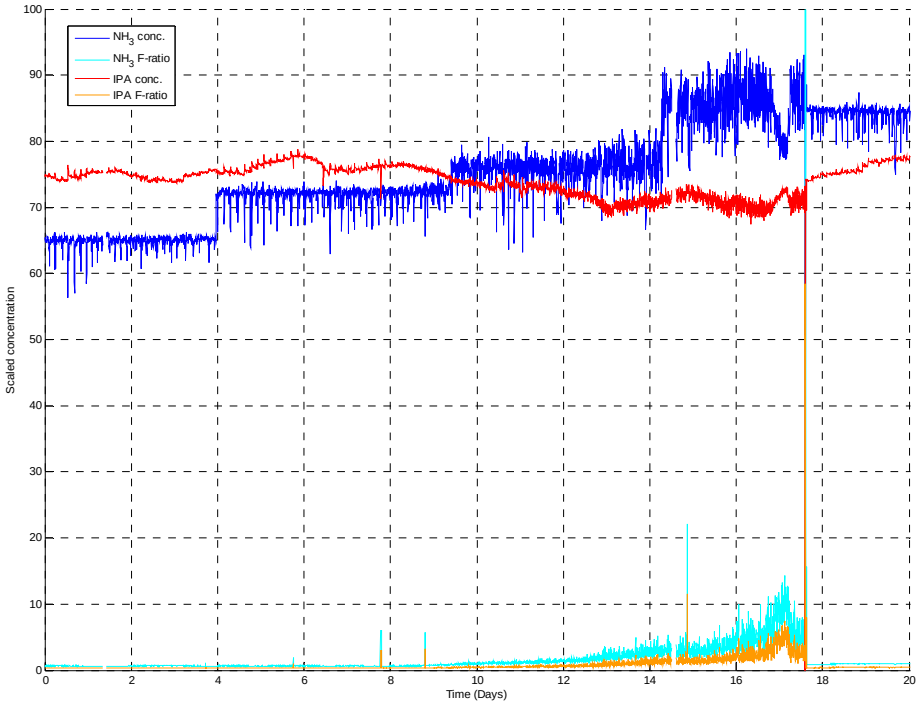


Figure 55 Deteriorating F-values over time of the on-line determination of ammonia and IPA lead to lower precision in the NIR predicted determination. The reason was identified as a malfunctioning valve, which was replaced on day 17.

The reason was identified as a malfunctioning valve in the on-line sampling system itself (as seen on Figure 1 in **Paper I**). It can obviously be argued that if no on-line system was in place, no valve could be malfunctioning. But the principle still applies and the cause of variation could very well be related elsewhere than the sampling system itself.

Recovering production from breakdowns

Even though highly automated systems always imply highly complex systems it can also prove beneficial during major breakdowns. September 24th 2003, the eastern part of Denmark suffered from an unusually large and extensive power failure for more than 6½ hours. The implications of such a long power failure in a pectin factory can be serious if pectin starts to gel in the production equipment, pipes, tanks and so on, because of cooling or lack of agitation, etc. For natural reasons, it is complex to restart an entire factory that was suddenly switched off rather than shut down in a controlled manner. But in that case, automation can be beneficial as it (once properly

started) it can provide measurements and assume control and relieve the operators to pursue other duties.
During the power failure, the ammonia dosing valve did not automatically close, leading to massive dosing of ammonia in the main tank. This is seen immediately after the system is put on-line in Figure 56 as a sharp increase in the predicted ammonia concentration. As the sensor is on-line rather than in-line directly in the main tank, a small dead volume of ammonia will need to be discarded before the actual amidation liquid reflecting the content in the main tank reaches the sensor.

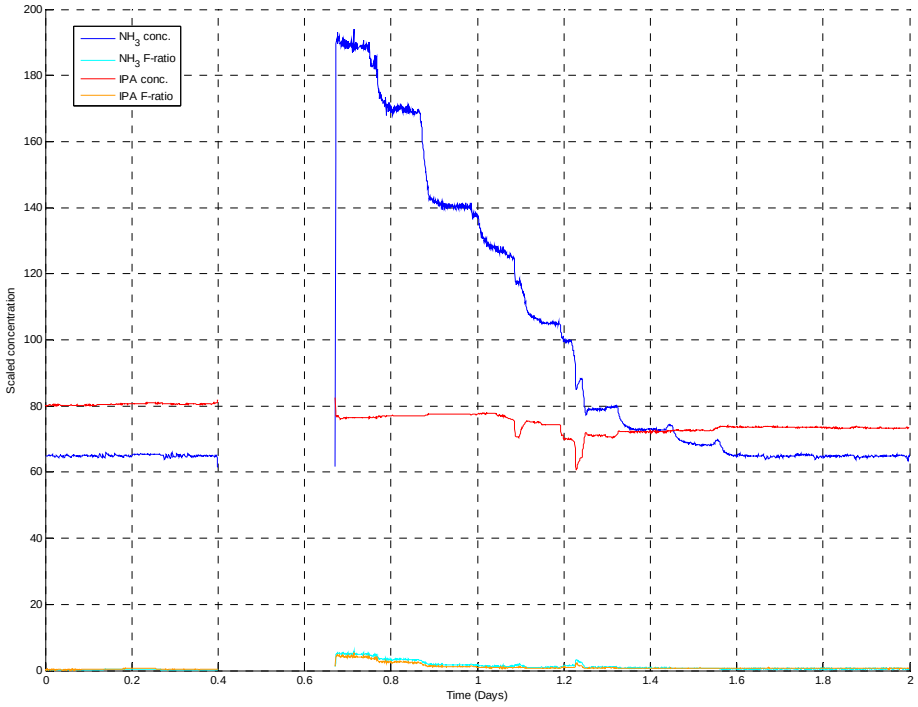


Figure 56 Scaled ammonia concentration before and after a 6½ hour power faliure affecting the factory.

It took more than half a day for the ammonia concentration to reach the desired set point. During that period an increase is seen in the F-ratio, warning that the spectral residuals are high as the ammonia concentration is above the calibrated window. The predicted ammonia concentrations are therefore a result of extrapolation.

Challenges in implementing PAT – Playing the PAT symphony

There are numerous challenges in implementing PAT in an operational form in a company. In a column in the magazine Spectroscopy Europe, it has been suggested that PAT is an abbreviation for "Possible Analytical Theory" rather than "Process Analytical Technology" (Davies, 2004). In order to be fully operational, PAT requires a fully orchestrated ensemble playing in tune. This is why introducing PAT can be compared to setting up a symphonic concert.

The musicians

PAT is a truly interdisciplinary field, and in order to successfully implement it, there is a need for solid teamwork. A multitude of skills is necessary: Operators, chemical and process engineers, chemometricians and statisticians, chemists, laboratory technicians, mechanical and automation engineers, the blue collar brigade (smiths, electricians and automation technicians), quality control and assurance professionals, and informatics and control experts, just to name a few who will be likely to find themselves involved. General knowledge of chemical unit operations, specific product and process knowledge, skills in data base mining, signal processing, multivariate data analysis and how to make structured input to and output from process data bases are required expertises. Obviously, there will be no music in time if there is no conductor (be it cracking the whip or swinging the baton). Management support on all levels is essential to focus on the overall goals rather than individuals or single department sub-goals.

The instruments

The instruments are not only the NIR instruments as some may believe, but all tools required to do the job. The categories of tools were mentioned in the introductory chapter such as multivariate tools for design, data acquisition and analysis, process analyzers, process control tools and the continuous improvement and knowledge management tools. In fact all the tools the musicians mentioned above are required to perform.

The sheets of music

Notes are the language of music. Notes are a standardized way of preserving and communicating music. Information needs to flow unhindered without the use of interpreters. Unfortunately, this is rarer than often seen in PAT solutions. Figure 57 shows the logistics, software and responsible companies involved in the on-line ammonia PAT implementation at CP Kelco.

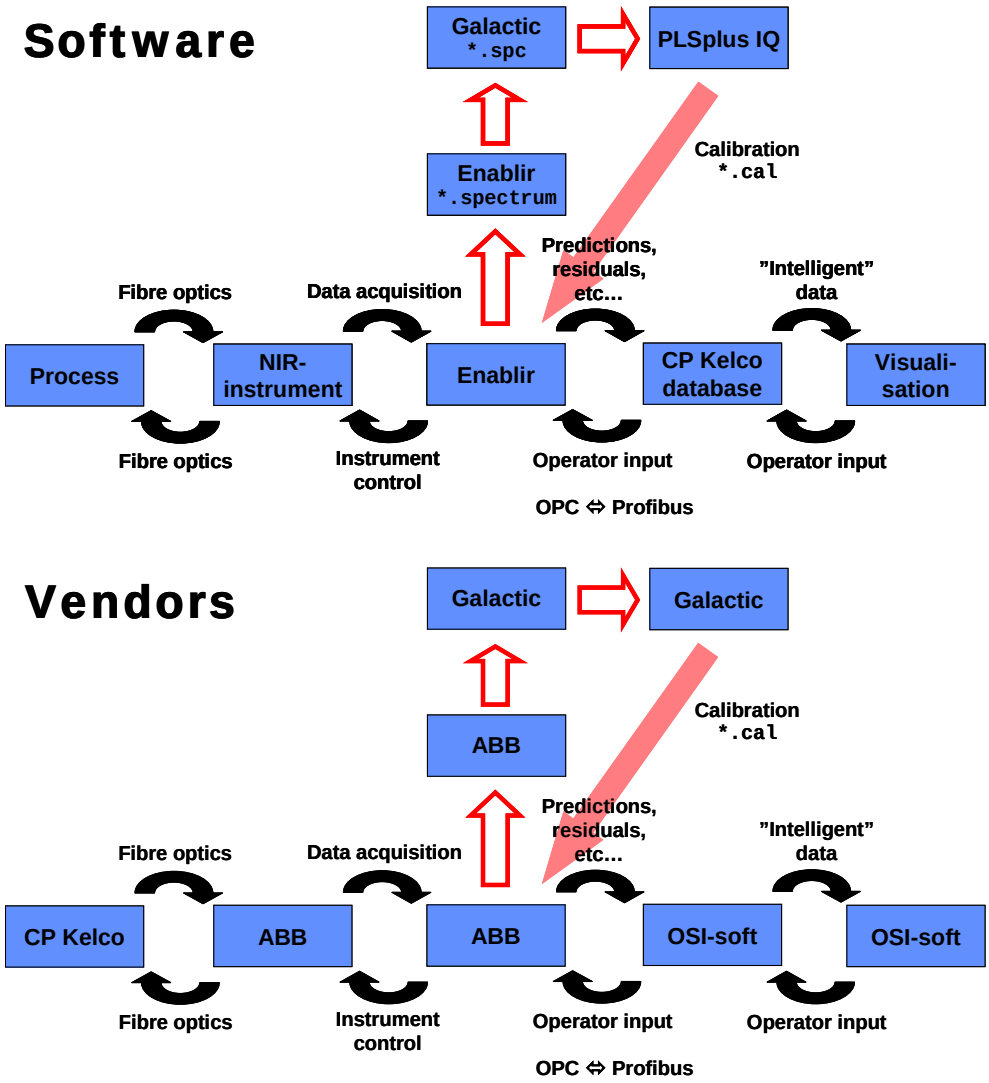


Figure 57 Overview of the logistics and software packages involved (top) and the responsible companies involved (bottom) in a PAT application. See the text for further details.

The "Process", which in this case is the on-line sample cells, appears in the lower left side of the top part figure. It is connected to the NIR instrument by fibre optics. The NIR instrument is controlled by the software package "Enablir" or FTSW100 as it is also called. Enablir translates the acquired NIR spectra to predictions which are fed to CP Kelcos process databases. As an extra challenge, Enablir sold by ABB does not support the "Profibus" field bus standard implemented at CP Kelco sold by Siemens, a competitor of

ABB. When converted, the data can then in turn be extracted from the process data base and visualized. This is the *on-line* loop. In the *off-line* loop lies the calibration maintenance. Enablir stores the NIR spectra in a proprietary file format called *.spectrum files. These have to be converted to another proprietary file format called *.spc files, before they can be read into a chemometric software package called PLS/IQ (virtually not updated since 1997) which can write the calibration files (also in a proprietary file format) that Enablir can read and process. If use of more modern chemometric software is desired, additional conversions back and forth are necessary. The bottom half of Figure 57 shows the companies responsible for the individual processes – clearly some interfaces to define.

As suggested above there is not a free flow of information. There is a need for universal instrument drivers, the use of non-proprietary file formats, a seamless exchange of data in data bases and the ability to handle spectral data in data bases, and to develop chemometric models on different platforms. Companies like Siemens/Profibus, ABB, Galactic, OSI soft, Umetrics and Mathworks (to name a few) need to define open standards and customers should demand them.

5. Conclusions

The research presented in this thesis has revealed a deeper insight and understanding of commercial pectin processing. Applications have been developed to measure ammonia and Isopropyl alcohol, which are identified as critical process parameters in the pectin production. The (near) real-time determination of the process parameters enables closed-loop control of the production systems. Furthermore, in-line, at-line and off-line applications that measure critical pectin quality attributes, such as the degree of esterification (DE) and its distribution and the degree of amidation (DA), have been devised and implemented. Feedback control of these parameters has created new value through improved product quality, quantity and processing capability. Specifically, raw material grading and selection, extraction, amidation and process optimisation have been addressed.

Spectroscopy

Two FT-NIR process analyzers with multiplexing capabilities, in total seven in-line, on-line and at-line measurement points have been implemented in the pectin production. Both transmittance and remote diffuse reflectance spectroscopy are employed, pushing the instruments, detectors, software, computers, probes and vendors to the edge of their capabilities – while at the same time pushing their edges. Furthermore, one off-line FT-NIR instrument for routine QC analysis and one off-line FT-NIR/IR instrument for method development have been supported.

For all applications correct sampling is crucial. New sampling protocols and devices have been developed. It is not always feasible, possible or practical to do everything flawlessly, but estimation of the sampling errors is imperative; both the unavoidable and also the avoidable sampling errors. Only then can fair compromises between how things *should* be done, and how they *are* done, be made.

The spectrometers have proven to be an eye into the production detecting normal process conditions most of the time fortunately. They have also been in place during process upsets providing reliable data far beyond normal

process conditions. This proves that spectroscopy and chemometrics can be robust and extrapolate when process conditions are out of specifications. This is an important aspect giving the applications credibility and allowing decision makers to have confidence in the numbers and focus on getting things right. Unfortunately, the applications have also been the cause of process upsets. Often, with a few notable exceptions, the instruments themselves have not been the cause of the problems, though they have taken most of the blame. It is usually the infrastructure around the instruments, primarily the software and operating systems, but also hardware, network and viruses that have grinded the applications to a halt. The sampling points themselves have failed as well as the interface to the control systems of the factory. It would be useful to have a pectin factory simulator as a test environment, before implementing new equipment, but this fails to be a stock item. Therefore, the hours one spends on foreseeing potential problems that may arise, figuring out how to counter them, and building logics and safeguards into systems, are hours well spent.
For the CAD analysis, a new generation of small and rather inexpensive UV-VIS spectrometers has been introduced. Based on a prism and diode array detector technology, they have no moving parts, expensive stepper motors, advanced optics or gratings to align and adjust. Their resolution is very high and while sensitivity is not the best compared to bench-top instruments, they certainly can do the job if the application allows sufficient light to get to the detector array. This type of instrument was interfaced in a novel application developed to estimate the deesterification pattern by measuring a dilution gradient of pectin in the presence of a dye with different affinity for acid groups, depending on whether they have ester groups or other acid groups as nearest neighbours. The setup is generic and can readily be used for other dye binding studies.

Chemometrics

Experimental design and multivariate data analysis is the cornerstone for extracting information from not only chemical problems, but a wide variety of investigations. Chemometrics can extract information from vast amounts of data and present them in a graphical format that most can perceive and relate to. Without chemometrics this study would not be feasible.
With chemometrics and experimental design, the significant can be separated from the insignificant. Methods have been devised to eliminate spectral artefacts, discard insignificant, redundant or noise-filled variables, and

detect outliers. A careful choice of spectral and chemometric pre-processing and variable selection eliminated the influence of temperature on the transmittance spectra of the amidation liquid. This removed the need for thermostatic control of the sampling points, which would have been expensive, not only to implement, but also to operate. In the thesis, an approach combining interval PLS with an exhaustive search of all one-region intervals in a spectral domain has been presented with the on-line ammonia data set used as an example. By grouping neighbouring variables into intervals of reasonable sizes, the computation time can be reduced appropriately.
Multi-way chemometrics has been applied to industrial data in order to resolve the dilution profiles generated in the CAD analysis. Common spectral profiles were identified for 33 samples with different DE and intermolecular distribution of ester groups and unique concentration profiles were resolved for the individual samples. Both PARAFAC2 and Multivariate Curve Resolution (MCR) could successfully resolve different components depending on the deesterification pattern induced in the pectins. An equation for estimating concentration profiles directly from a PARAFAC2 model has been proposed. The improved flexibility of MCR as compared with PARAFAC2 improves the fit of the resolved concentration profiles of MCR to the acquired data. Information with regard to amount and dilution time resolved in the concentration profiles can be related to the deesterification pattern of the pectins, and models quantifying the amount of random and blocky acid groups using the CAD approach have been implemented at CP Kelco. The resolved concentration profiles enhance the understanding of the system, but unfold PLS on the whole sample matrix was equally efficient in predicting the deesterification pattern. However, with an added risk of overfitting in the unfolded PLS models due to an inherent higher number of variables.

PAT in pectin production

PAT has *not* been implemented at CP Kelco as an overall strategy – it has been exemplified and evaluated by the research carried out during this PhD study. In 2003, CP Kelco was recognized with Siemens Technology Services for implementing a large automation solution based on a common Fieldbus. Virtually all sensors, classical as well as "new" multivariate sensors, are connected to the PI process database solution from OSI-soft. Thus, the infrastructure at CP Kelco is in place, and PAT is being done every day.
In that regard, a NIR instrument is just another sensor, albeit more expensive and more complex than the average pH electrode, but just like a pH

electrodes needs calibration and maintenance, so does a NIR instrument. And that may initially be a bit more elaborate and resource demanding than the pH electrode.
On the other hand, a NIR instrument is a highly versatile sensor, capable of measuring things most other sensors fail to see. It is crucial to have the vision and dare go where others have backed off and put these "ambitions" sensors on-line. On top of implementing a NIR based system providing "just" accurate, precise, stable, fast and reliable predictions of ammonia far beyond the initial expectation, much has been learned about the pectin process. Quality has been built in by design rather than by post-process testing in QC. Not just a quick responding, reproducible and reliable ammonia dosing system has been implemented removing a tedious task of titrating process samples once an hour around the clock. Quicker turnovers from one pectin recipe to another are now a reality. This makes production planning more flexible, a great benefit for any supply chain organisation. Smaller changes in ammonia concentration can reliably be made enabling further optimization of the yield without compromising the desired pectin quality. As a spin-off, the NIR instrument was capable of measuring the Isopropyl alcohol (IPA) content in the amidation liquid as well. As a consequence a process parameter regarded as not critical and considered within control was, as a result of timely and reliable measurements around the clock, identified as critical for the process and hereto new process relations were suddenly grasped. This further enabled the elimination of unwanted process variation and provided new knowledge of the pectin production.
The in-line and at-line DE/DA measurement on wet pectin press cake has been implemented by using optical fibre probes and remote diffuse reflectance spectroscopy. The measurement to measurement variability is high, but the moving average of the measurements is accurate compared with final QC measurements on the dry pectin powder. This illustrates the benefit when many measurements are made and a large amount of material is sampled. The CAD application is not a PAT system as such, but it has been designed to go at-line if desired. It requires minimal sample preparation, is highly automated, and the analysis time per sample is low enabling a high possible sample throughput.

6. Perspectives

PAT as a tool will impact all industries in the 21st century. More and more sensors will be interfaced into production environments and more and more data will automatically be harvested into large process data bases. Companies which are better at turning data into information and decisions, will prevail. Turning data into useful information requires the data to be digested. Chemometrics is crucial in this process. Once the data has been turned into information that information will need to be displayed in a user friendly manner for the decision makers at all levels: Turning the information into knowledge! Decisions will to a higher degree have to be made "on-line" as things are happening rather than "off-line" on the basis of carefully prepared reports analysing the process in retrospect, as it once was.

There will be a need for improved understanding of the value, process and quality chain so the optimal process parameters can be chosen for different quality and batches of raw materials. The identification of critical process parameters and their influence on product attributes and product functionalities are important. Furthermore, companies which are better at relating their product functionalities to their customers' individual applications will be winners.

There will be a need for educating process chemometricians and upgrading current staff. The Quality and Technology group at University of Copenhagen has started a two-year MSc programme in Process Analytical Technology where the graduates obtain practical and theoretical competences within experimental design, multivariate data analysis, on-line measurement systems and statistical process analysis.

Multi-way chemometrics will continue to become more and more important. Not only hyphenated analytical instruments will be more common and inexpensive, but also batch analysis will be the standard of tomorrow, where entire batches are seen as a sample with local batch time and process variables (perhaps multi-way?) constitute the other modes. The benefit will be a more unified understanding of the process.

References

(ABB, 2001) *FT-IR reference manual*, Revision 1.1, August 2001, ABB Inc.

(Andersson, 1999) C.A. Andersson, Direct orthogonalization, *Chemometrics and Intelligent Laboratory Systems*, **47** (1999), 51-63

(ASDI, 2007) http://www.asdi.com/nir-chart_grid_rev-3.pdf as of June 2007

(Axelos, 1991) M.A.V. Axelos and J.-F. Thibault, The chemistry of low-methoxyl pectin gelation, in *The chemistry and technology of pectins* (Eds. R. Walter and S. Taylor), 1st Edition, Academic Press, New York, 1991, 109-118

(Beebe, 1998) K.R. Beebe, R.J. Pell and M.B. Seasholtz, *Chemometrics: A practical guide*, Wiley, New York, 1998

(Blaser, 1995) W.W. Blaser, R.A. Bredeweg, R.S. Harner, M.A. LaPack, A. Leugers, D.P. Martin, R.J. Pell, J. Workman Jr. and L.G. Wright, Process analytical chemistry, *Analytical Chemistry*, **67** (1995), 47R-70R

(Booksh, 1994) K.S. Booksh and B.R. Kowalski, Theory of analytical chemistry, *Analytical Chemistry*, **66** (1994), 782A-791A

(Braccini, 2001) I. Braccini and S. Pérez, Molecular basis of Ca^{2+}-induced gelation in alginates and pectins: The egg-box model revisited, *Biomacromolecules*, **2** (2001), 1089-1096

(Brejnholt, 2007) S.M. Brejnholt, Pectin, submitted to appear in *Food stabilisers, thickeners and gelling agents* (Ed. A. Imerson), Blackwell, Oxford, 2007

(Bro, 1996a) R. Bro, Multiway calibration. Multilinear PLS, *Journal of Chemometrics*, **10** (1996), 47-61

(Bro, 1996b) R. Bro, *Håndbog i multivariabel kalibrering*, Jordbrugsforlaget, Copenhagen, 1996

(Bro, 1997) R. Bro, PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, **38** (1997), 149-171

(Bro, 1998) R. Bro, *Multi-way analysis in the food industry – Models, algorithms and applications*, Doctoral Thesis, University of Amsterdam, Amsterdam, 1998

(Bro, 1999) R. Bro, C.A. Andersson and H.A.L. Kiers, PARAFAC2 – Part II. Modeling chromatographic data with retention time shifts, *Journal of Chemometrics*, **13** (1999), 295-309

(Bro, 2003) R. Bro and A.K. Smilde, Centering and scaling in component analysis, *Journal of Chemometrics*, **17** (2003), 16-33

(Bryant, 1949) E.F. Bryant, *Acid amides of pectinic acid and process for their preparation*, U.S. Patent, 2,480,710, August 30th 1949

(Callis, 1987) J.B. Callis, D.L. Illman and B.R. Kowalski, Process analytical chemistry, *Analytical Chemistry*, **59** (1987), 624A-637A

(Carpita, 1993) N.C. Carpita and D. Gibeaut, Structural models of primary cell walls in flowering plants: Consistency of molecular structure with the physical properties of the walls during growth, *The Plant Journal*, **3** (1993), 1-30

(Carroll, 1970) J.D. Carroll and J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalisation of "Eckart-Young" decomposition, *Psychometrika*, **35** (1970), 283-319

(CP Kelco, 2003) *Genu® Pectin Book*, rev. 06/04, 2003

(Daas, 1999) P.J.H. Daas, K. Meyer-Hansen, H.A. Schols, G.A. de Ruiter and A.G.J. Voragen, Investigation of the non-esterified galacturonic acid distribution in pectin with endopolygalacturonase, *Carbohydrate Research*, **318** (1999), 135-145

(Daas, 2000) P.J.H. Daas, A.G.J. Voragen and H.A. Schols, Investigation of the galacturonic acid distribution of pectin with enzymes part 2 – Characterization of non-esterfied galacturonic acid sequences in pectin with endo-polygalacturonase, *Carbohydrate Research*, **326** (2000), 120-129

(Davies, 2004) A.M.C. Davies, What is PAT?, *Spectroscopy Europe*, **16** (2004), 33-34

(Dhanoa, 1994) M.S. Dhanoa, S.J. Lister, R. Sanderson and R.J. Barnes, The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra, *Journal of Near Infrared Spectroscopy*, **2** (1994), 43-47

(Douglas, 1913) R. Douglas, *Food product*, U.S. Patent 1,082,682, December 30th 1913

(Engelsen, 1996) S.B. Engelsen and L. Nørgaard, Comparative vibrational spectroscopy for determination of quality parameters in amidated pectins as evaluated by chemometrics, *Carbohydrate Polymers*, **30** (1996), 9-24

(Esbensen, 2000) K.H. Esbensen, *Multivariate data analysis – In practice*, 4th Edition, Camo, Oslo, 2000

(Escandar, 2006) G.M. Escandar, P.C. Damiani, H.C. Goicoechea and A.C. Olivieri, A review of multivariate calibration methods applied to biomedical analysis, *Microchemical Journal*, **82** (2006), 29-42

(FAO, 2001) Compendium of food additives specifications/add. 9, *FAO food and nutrition paper no. 52*, Rome, Italy, 2001

(FAO, 2007) http://faostat.fao.org/site/336/DesktopDefault.aspx?PageID=336 as of June 2007

(FDA, 2004) United States Food and Drug Administration, *Guidance for industry PAT – A framework for innovative pharmaceutical development, manufacturing and quality assurance*, U.S. Department of Health and Human Services, Rockville, 2004

(Geladi, 1985) P. Geladi, D. MacDougall and H. Martens, Linearization and scatter-correction for near-infrared reflectance spectra of meat, *Applied Spectroscopy*, **39** (1985), 491-500

(Geladi, 1987) P. Geladi, Analysis of multi-way (multi-mode) data, *Chemometrics and Intelligent Laboratory Systems*, **7** (1987), 11-30

(Gerlach, 2003) R.W. Gerlach and J.M. Nocerino, *Guidance for obtaining representative laboratory analytical subsamples from particulate laboratory samples*, EPA 600-R-03-027, U.S. Environmental Protection Agency, Washington, DC, 2003

(Gnanasambandam, 2000) R. Gnanasambandam and A. Proctor, Determination of pectin degree of esterification by diffuse reflectance Fourier transform infrared spectroscopy, *Food Chemistry*, **68** (2000), 327-332

(Grant, 1973) G.T. Grant, E.R. Morris, D.A. Rees, P.J.C. Smith and D. Thom, Biological interactions between polysaccharides and divalent cations: The egg box model. *FEBS Letters*, **32** (1973), 195-198

(Grasdalen, 1988) H. Grasdalen, O.E. Bakøy and B. Larsen, Determination of the degree of esterification and the distribution of methylated and free carboxyl groups in pectins by ^{1}H-NMR spectroscopy, *Carbohydrate Research*, **184** (1999), 183-191

(Guillotin, 2005) S.E. Guillotin, E.J. Bakx,. P. Boulenguer, J. Mazoyer, H.A. Schols and A.G.J. Voragen, Populations having different GalA blocks characteristics are present in commercial pectins which are chemically similar but have different functionalities, *Carbohydrate Polymers*, **60** (2005), 391-398

(Guillotin, 2006) S.E. Guillotin, J. van Kampen, P. Boulenguer, H.A. Schols and A.G.J. Voragen, Degree of blockiness of amide groups as indicator for difference in physical behavior of amidated pectins, *Biopolymers*, **82** (2006), 29-37

(Guillotin, 2007) S.E. Guillotin, E.J. Bakx, P. Boulenguer, H.A. Schols and A.G.J. Voragen, Determination of the degree of substitution, degree of amidation and degree of blockiness of commercial pectins by using capillary electrophoresis, *Food Hydrocolloids*, **21** (2007), 444-451

(Gy, 1998) P. Gy, *Sampling for analytical purposes*, Wiley, New York, NY, 1998

(Haaland, 1988) D.M. Haaland and E.V. Thomas, Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information, *Analytical Chemistry*, **60** (1988), 1193-1202

(Harshman, 1970) R. Harshman, Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis, *UCLA Working Papers in Phonetics*, **16** (1970), 1-84

(Hassell, 1998) D.C. Hassell and E.M. Bowman, Process analytical chemistry for spectroscopists, *Applied Spectroscopy*, **52** (1998), 18A-29A

(Hussain, 2002) A.S. Hussain, *The subcommittee on process analytical technologies (PAT): Overview and objectives*, FDA's Advisory Committee for Pharmaceutical Science, Panel Meeting in Gaithersburg, MD, February 25, 2002

(IPPA, 2007) http://www.ippa.info/history_of_pectin.htm as of June 2007

(Ishii, 1979) S. Ishii, K. Kiho, S. Sugiyama and H. Sugimoto, Low-methoxyl pectin prepared by pectinesterase from *Aspergillus japonicus*, *Journal of Food Science*, **44** (1979), 611-614

(Jackson, 1991) J.E. Jackson, *A user's guide to principal components*, Wiley, New York, 1991

(Jaumot, 2005) J. Jaumot, R. Gargallo, A. de Juan and R. Tauler, A graphical user-friendly interface for MCR-ALS: A new tool for multivariate curve resolution in MATLAB, *Chemometrics and Intelligent Laboratory Systems*, **76** (2005), 101-110

(Jiang, 2001) C.-M. Jiang, M.-C. Wu, W.-H. Chang and H.-M. Chang, Determination of random- and blockwise-type de-esterified pectins by capillary zone electrophoresis, *Journal of Agricultural and Food Chemistry*, **49** (2001), 5584-5588

(Joseph, 1949) G.H. Joseph, A.H. Kieser and E.F. Bryant, High-polymer, ammonia-demethylated pectinates and their gelation, *Food Technology*, **3** (1949), 85-90

(de Juan, 2003) A. de Juan and R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures revisiting latest trends in multivariate resolution, *Analytica Chimica Acta*, **500** (2003), 195-210

(Keller, 1992) H.R. Keller and D.L. Massart, Evolving factor analysis. *Chemometrics and Intelligent Laboratory Systems*, **12** (1992), 209-224

(Kiers, 1999) H.A.L. Kiers, J.M.F. ten Berge and R. Bro, PARAFAC2 – Part I. A direct fitting algorithm for the PARAFAC2 model, *Journal of Chemometrics*, **13** (1999), 275-294

(Kim, 1978) W.J. Kim, C.B.J. Smit, V.N.M. Rao, Demethylation of pectin using acid and ammonia, *Journal of Food Science*, **43** (1978), 74-78

(Kohn, 1968) R. Kohn, I. Furda and Z. Kopec, Distribution of free carboxyl groups in pectin molecule after treatment with pectin esterase, *Collection of Czechoslovak Chemical Communications*, **33** (1968), 264-269

(Kravtchenko, 1992) T.P. Kravtchenko, A.G.J. Voragen and W. Pilnik, Analytical comparison of three industrial pectin preparations, *Carbohydrate Polymers*, **18** (1992), 17-25

(Leardi, 2001) R. Leardi, Genetic algorithms in chemometrics and chemistry: A review, *Journal of Chemometrics*, **15** (2001), 559-569

(Levy, 1988) G.C. Levy, How changes in computer technology are revolutionizing the practice of chemistry, *Journal of Chemical Information and Computer Science*, **28** (1988), 167-174

(Limberg, 2000) G. Limberg, R. Korner, H.C. Buchholt, T.M.I.E. Christensen, P. Roepstorff and J.D. Mikkelsen, Analysis of pectin structure part 3 – Quantification of the amount of galacturonic acid residues in blocksequences in pectin homogalacturonan by enzymatic fingerprinting with exo- and endo-polygalacturonase II from *Aspergillus niger*, *Carbohydrate Research*, **327** (2000), 321-332

(Lorber, 1997) A. Lorber, K. Faber and B.R. Kowalski, Net analyte signal calculation in multivariate calibration, *Analytical Chemistry*, **69** (1997), 1620-1626

(Louwerse, 1999) D.J. Louwerse, A.K. Smilde and H.A.L. Kiers, Cross-validation of multiway component models, *Journal of Chemometrics*, **13** (1999), 491-510

(Madden, 2000) D. Madden, In a jam and out of juice, NCBE / Unilever educational guide, Reading, 2000, http://www.ncbe.reading.ac.uk/NCBE/PROTOCOLS/juice.html as of June 2007

(Maeder, 1986) M. Maeder and A.D. Zuberbühler, The resolution of overlapping chromatographic peaks by evolving factor analysis, *Analytica Chimica Acta*, **181** (1986), 287-291

(Maeder, 1987) M. Maeder, Evolving factor analysis for the resolution of overlapping chromatographic peaks, *Analytical Chemistry*, **59** (1987), 527-530

(Martens, 1989) H. Martens and T. Næs, *Multivariate calibration*, Wiley, Chichester, 1989

(Martens, 1998) H.A. Martens and P. Dardenne, Validation and verification of regression in small data sets, *Chemometrics and Intelligent Laboratory Systems*, **44** (1998), 99-121

(Martens, 2003) H. Martens, J.P. Nielsen and S.B. Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures, *Analytical Chemistry*, **75** (2003), 394-404

(McCann, 1991) M.C. McCann and K. Roberts, Architecture of the primary cell wall, in *The Cytoskeletal Basis of Plant Growth and Form* (Ed. C.W. Lloyd), Academic Press, London, 1991

(McCready, 1954) R.M. McCready and C.G. Seegmiller, Action of pectic enzymes on oligogalacturonic acids and some of their derivatives, *Archives of Biochemistry and Biophysics*, **50** (1954), 440-450

(McNaught, 1997) A.D. McNaught and A. Wilkinson, *Compendium of Chemical Terminology, The Gold Book*, 2nd Edition, Blackwell Science, Oxford, 1997

(Meyer, 2006) T. Meyer, J. Oelichmann and H. Kellerhals, Resolution and suppression of mechanical noise in FT–NIR spectroscopy, *Trends in Analytical Chemistry*, **25** (2006), 19-23

(Miller, 1971) L. Miller and J.D. MacMillan, Purification and pattern of action of pectinesterase fom *Fusarium oxysporum* f. sp. *vasinfectum*, *Biochemistry*, **10** (1971), 570-576

(Morris, 2007) V.J. Morris, Gels, in *The Chemical Physics of Food* (Ed. P. Belton), 1st Edition, Blackwell Publishing, Oxford, 2007

(Munck, 1998) L. Munck, L. Nørgaard, S.B. Engelsen, R. Bro and C.A. Andersson, Chemometrics in food science – A demonstration of a highly exploratory, inductive evaluation strategy of fundamental scientific significance, *Chemometrics and Intelligent Laboratory Systems*, **44** (1998), 31-60

(Munck, 2006) L. Munck, *The revolutionary aspect of chemometric technology*, Samfundslitteratur, Copenhagen, 2006

(Needs, 2001) P.W. Needs, N.M. Rigby, S.G. Ring and A.J. MacDougall, Specific degradation of pectins via a carbodiimide-mediated Lossen rearrangement of methyl esterified galacturonic acid residues, *Carbohydrate Research*, **333** (2001), 47-58

(Nørgaard, 2000) L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, Interval partial least squares regression (iPLS): A comparative study with an example from near-infrared spectroscopy, *Applied Spectroscopy*, **54** (2000), 413-419

(Pavia, 2001) D.L. Pavia, G.M. Lampman and G.S. Kriz, *Introduction to spectroscopy*, 3rd Edition, Harcourt, Orlando, 2001

(Pérez, 2003) S. Pérez, M.A. Rodríguez-Carvajal and T. Doco, A complex plant cell wall polysaccharide: Rhamnogalacturonan II. A structure in quest of a function, *Biochemie*, **85** (2003), 109-121

(Petersen, 2005) L. Petersen and K.H. Esbensen, Representative process sampling for reliable data analysis – A tutorial, *Journal of Chemometrics*, **19** (2005), 625-647

(Potter, 2005) B.B. Potter and J.C. Wimsatt, *Determination of total organic carbon and specific UV absorbance at 254 nm in source water and drinking water*, EPA 600-R-05-055, U.S. Environmental Protection Agency, Cincinnati, Ohio, 2005

(Prazen, 1998) B.J. Prazen, R.E. Synovec and B.R. Kowalski, Standardization of second-order chromatographic/spectroscopic data for optimum chemical analysis, *Analytical Chemistry*, **70** (1998), 218-225

(Ridley, 2001) B.L. Ridley, M.A. O'Neill and D. Mohnen, Pectins: Structure, biosynthesis, and oligogalactoronide-related signalling, *Phytochemistry*, **57** (2001), 929-967

(Rolin, 1990) C. Rolin and J.A. de Vries, in *Food gels* (Ed. P. Harris), Elsevier, Amsterdam, 1990, 401-434

(Rolin, 2002) C. Rolin, Commercial pectin preparations, in *Pectins and their manipulation* (Eds. G.B. Saymour and J.P. Knox), Blackwell Publishing, Oxford, 2002, 222-241

(Rosenbohm, 2003) C. Rosenbohm, I. Lundt, T.M.I.E. Christensen, N.W.G. Young, Chemically methylated and reduced pectins: Preparation, characterisation by ^{1}H NMR spectroscopy, enzymatic degradation, and gelling properties, *Carbohydrate Research*, **338** (2003), 637-649.

(Růžička, 1988) J. Růžička and E.H. Hansen, *Flow injection analysis*, 2nd Edition, Wiley, New York, 1988

(Savary 2003) B.J. Savary, A.T. Hotchkiss, M.L. Fishman, R.G. Cameron and R.G. Shatters, Development of a valencia orange pectin methyl esterase for generating novel pectin products, in *Advances in pectin and pectinase research* (Eds. A.G.J. Voragen, H.A. Schols and R.G.F. Visser), Kluwer, Amsterdam, 2003, 345-361

(Savitzky, 1964) A. Savitzky and M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, **36** (1964), 1627-1639

(Schols, 2002) H.A. Schols and A.G.J. Voragen, The chemical structure of pectins, in *Pectins and their manipulation* (Eds. G.B. Saymour and J.P. Knox), Blackwell Publishing, Oxford, 2002, 1-29

(Séné, 1994) C.F.B. Séné, M.C. McCann, R.H. Wilson and R. Grinter, Fourier-transform Raman and Fourier-transform infrared spectroscopy: An investigation of five higher plant cell walls and their components. *Plant Physiology*, **106** (1994), 1623-1630

(Silverstein, 1991) R.M. Silverstein, G.C. Bassler and T.T. Morrill, *Spectrometric identification of organic compounds*, 5th Edition, Wiley, New York, 1991

(Sjöström, 1983) M. Sjöström, S. Wold, W. Lindberg, J. A. Persson and H. Martens, A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables, *Analytica Chimica Acta*, **150** (1983), 61-70.

(Smilde, 2004) A.K. Smilde, R. Bro and P. Geladi, *Multi-way analysis with applications in the chemical sciences*, John Wiley & Sons, Chichester, 2004

(Solms, 1955) J. Solms and H. Deuel, Über den mechanismus der enzymatischen verseifung von pektinstoffen, *Helvetica Chimica Acta*, **38** (1955), 321-329

(Speiser, 1945) R. Speiser, C.H. Hills and C.R. Eddy, The acid behavior of pectinic acids, *The Journal of Physical Chemistry*, **49** (1945), 328-343

(Speiser, 1946) R. Speiser and C.R. Eddy, Effect of molecular weight and method of deesterification on the gelling behavior of pectin, *Journal of the American Chemical Society*, **68** (1946), 287-293

(SURA, 2007) http://www.sura.org/2000/SURA_Electromagnetic_Spectrum_Full_Chart.jpg as of June 2007

(Synytsya, 2003) A. Synytsya, J. Čopíková, P. Matějka and V. Machovič, Fourier transform Raman and infrared spectroscopy of pectins, *Carbohydrate Polymers*, **54** (2003), 97-106

(Tauler, 1995a) R. Tauler, Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, **30** (1995), 133-146

(Tauler, 1995b) R. Tauler, A.K. Smilde and B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *Journal of Chemometrics*, **9** (1995), 31-58

(Thakur, 1997) B.R. Thakur, R.K. Singh and A.K. Handa, Chemistry and uses of pectin – A review, *Critical Reviews in Food Science and Nutrition*, **37** (1997), 47-73

(Thennadil, 2006) S.N. Thennadil, H. Martens and A. Kohler, Physics-based multiplicative scatter correction approaches for improving the performance of calibration models, *Applied Spectroscopy*, **60** (2006), 315-321

(Thibault, 2003) J.-F. Thibault and M.-C. Ralet, Physico-chemical properties of pectins in the cell walls and after extraction, in *Advances in pectin and pectinase research* (Eds. A.G.J. Voragen, H.A. Schols and R.G.F. Visser), Kluwer, Amsterdam, 2003, 91-105

(Tuerena, 1982) C.E. Tuerena, A.J. Taylor and J.R. Mitchell, Evaluation of a method for determining the free carboxyl group distribution in pectins, *Carbohydrate Polymers*, **2** (1982), 193-203

(Umetrics, 2006), L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström and S. Wold, *Multi- and megavariate data analysis, Part I: Basic principles and applications*, Umetrics, Umeå, 2006

(Vidrine, 2000) D.W. Vidrine, Mid-infrared spectroscopy in chemical process analysis, in *Spectroscopy in process analysis* (Ed. J.M. Chalmers), Sheffield Academic Press, Sheffield, 2000, 96-138

(Vincken, 2003) J.P. Vincken, H.A Schols, R.J.F.J. Oomen, M.C. McCann, P. Ulvskov, A.G.J. Voragen and R.G.F. Visser, If homogalacturonan were a side chain of rhamnogalacturonan I. Implications for cell wall architecture. *Plant Physiology*, **132** (2003), 1781-1789

(de Vries, 1986) J.A. de Vries, M. Hansen, J. Søderberg, P.-E. Glahn and J.K. Pedersen, Distribution of methoxyl groups in pectins, *Carbohydrate Polymers*, **6** (1986), 165-176.

(Westerhuis, 1999) J.A. Westerhuis, T. Kourti and J.F. MacGregor, Comparing alternative approaches for multivariate statistical analysis of batch process data, *Journal of Chemometrics*, **13** (1999), 397-413

(Westerlund, 1991) E. Westerlund, P. Åman, R.E. Andersson and R. Andersson, Investigation of the distribution of methyl ester groups an pectin by high-field ^{13}C NMR, *Carbohydrate Polymers*, **14** (1991), 179-187

(Williams, 2005) P.A. Williams, C. Sayers, C. Viebke and C. Senan, Elucidation of the emulsification properties of sugar beet pectin, *Journal of Agricultural and Food Chemistry*, **53** (2005), 3592-3597

(Willats, 2001) W.G.T. Willats, L. McCartney, W. Mackie and J.P. Knox, Pectin: Cell biology and prospects for functional analysis, *Plant Molecular Biology*, **47** (2001), 9-27

(Willats, 2006) W.G.T. Willats, J.P. Knox and J.D. Mikkelsen, Pectin: New insights into an old polymer are starting to gel, *Trends in Food Science and Technology*, **17** (2006), 97-104

(Windig, 1991) W. Windig and J. Guilment, Interactive self-modeling mixture analysis, *Analytical Chemistry*, **59** (1991), 1425-1432

(Winning, 2007) H. Winning, N. Viereck, L. Nørgaard, J. Larsen and S.B. Engelsen, Quantification of the degree of blockiness in pectins using ^{1}H NMR spectroscopy and chemometrics, *Food Hydrocolloids*, **21** (2007), 256-266

(Wise, 2001) B.M. Wise, N.B. Gallagher and E.B. Martin, Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch, *Journal of Chemometrics*, **15** (2001), 285-298

(Wold, 1978) S. Wold, Cross validatory estimation of the number of components in factor and principal component models, *Technometrics*, **20** (1978), 397-406

(Wold, 1987) S. Wold, K.H. Esbensen and P. Geladi, Principal component analysis – A tutorial, *Chemometrics and Intelligent Laboratory Systems*, **2** (1987), 37-52

(Wold, 1998) S. Wold, H. Antti, F. Lindgren and J. Öhman, Orthogonal signal correction of near-infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, **44** (1998), 175-185

(Wold, 2001) S. Wold, M. Sjöström and L. Eriksson, PLS-regression: A basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58** (2001), 109-130

(Workman, 2005) J. Workman Jr., *Chemometrics and PAT: What does it all mean?*, http://www.spectroscopyonline.com, January 2005.

(Wülfert, 2000) F. Wülfert, W.T. Kok, O.E. de Noord and A.K. Smilde, Correction of temperature-induced spectral variation by continuous piecewise direct standardization, *Analytical Chemistry*, **72** (2000), 1639-1644

(Yapo, 2007) B.M Yapo, P. Lerouge, J.-F. Thibault and M.-C. Ralet, Pectins from citrus peel cell walls contain homogalacturonans homogenous with respect to molar mass, rhamnogalacturonan I and rhamnogalacturonan II, *Carbohydrate Polymers*, **69** (2007), 426-435

Poster I

Novel dilution pectin profiling by DAD UV-VIS and three-way analysis

Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Rasmus Bro, Søren Balling Engelsen

Presented at the 9th Conference on Chemometrics in Analytical Chemistry, Lisbon, 2004.

Novel dilution pectin profiling by DAD UV-VIS and three-way analysis

Christian B. Zachariassen[a,b,*], Jan Larsen[a], Frans van den Berg[b], Rasmus Bro[b], Søren B. Engelsen[b].

[a] CP Kelco, Quality Control Development Group, Ved Banen 16, DK-4623 Lille Skensved, Denmark

[b] The Royal Veterinary and Agricultural University, Department of Food Science, Food Technology / Centre for Advanced Food Studies, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

[*] Corresponding author. E-mail: christian.zachariassen@cpkelco.com, Telephone: +45 5616 5616

Background

Pectin mainly consists of a linear chain of poly-α-(1→4)-D-galacturonic acid with varying degree of esterification (%DE). The distribution of the (methyl) ester groups on the pectin carbohydrate backbone is an important structural parameter that affects pectin functionality in customer applications. In a study conducted at CP Kelco, the world's largest pectin producer, different pectins extracted from natural citrus peels are investigated. In order to span the sample set beyond natural variation, artificial samples of re-methylated pectins are enzymatically de-esterified by two different enzymes, de-esterifing the pectin in a "Random" or "Blocky" fashion.

The exact distribution of ester groups on a pectin backbone is a property that can only be defined for an individual pectin polymer, and this information is as such of no commercial interest. It is impossible to extract or produce identical pectin polymer molecules from natural sources. However, a method for measuring average distribution of ester groups in bulk pectin is highly desired. How "Blocky" or "Random" pectin is, is not easily expressed as a single number. The method will have to take into account that the average total degree of esterification (%DE) will vary from sample to sample.

Analysis

An exploratory study has been designed using a pectin binding dye, in which – during a controlled dilution gradient in a liquid system – a diode array detector (DAD) records UV-VIS spectra, yielding a three-way matrix of samples × wavelength × dilution time.

In total 12 pectins have been analysed. The natural pectin extracted from citrus peel (M0), 6 pectins de-esterified from M0 to varying %DE by an enzyme known to de-esterify in a blocky manner (B1-B6) and 5 pectins de-esterified from M0 to varying %DE by an enzyme known to de-esterify in a random manner (R1-R5).

The uncentered 2D spectral fingerprints have been analyzed with a PARAFAC2[1] algorithm[2] using non-negativity constraints imposed on the sample mode, with shifting loadings in the dilution time direction. The sample mode loadings from PARAFAC2 correlate reasonably well (r = 0.73) to the degree of esterification, although no information about the %DE is used in the PARAFAC2 algorithm. This is a good indication that the spectra plus dilution profiles contain relevant information about the pectin.

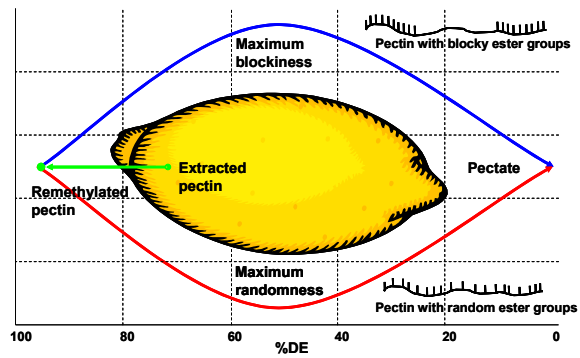
An uncentered multi-way PLS calibration[3] on the degree of esterification predicts %DE with a correlation of r = 0.95 using 4 PLS components. The score plot from the multi-way PLS calibration fits the initial problem definition.

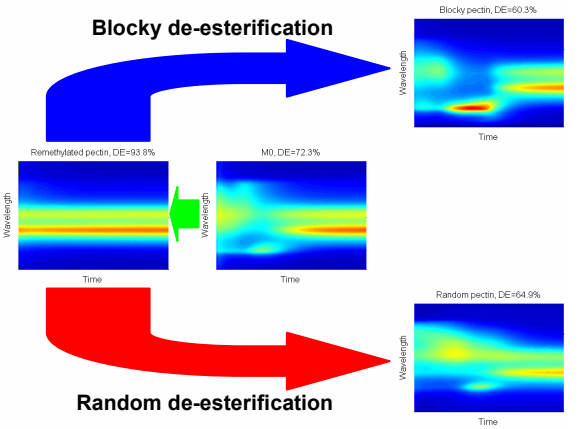
Conclusion

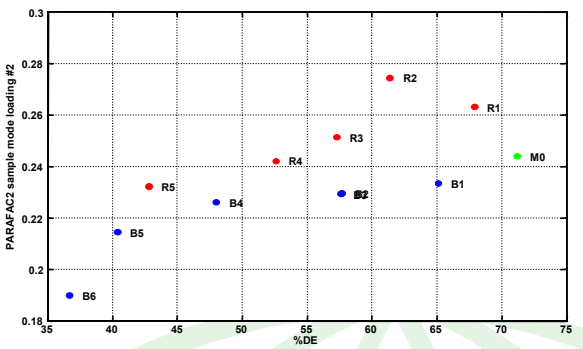
The very successful preliminary studies reveal that the use of multi-way PLS to correlate the spectral landscapes to the degree of esterification yields interpretable models. It resolves the dilution time dependent spectral profiles and recovers information about the average distribution of methyl ester groups so the average blockiness of a pectin sample can readily be assessed. PARAFAC2 scores also correlate to the degree of esterification yielding relevant information. Further work on an enhanced data set has already commenced, in order to better understand the spectral profiles in relation to the chemistry and kinetics of the system. Additional approaches to the spectral resolution e.g. algorithms based on Alternating Least Squares will also be investigated.

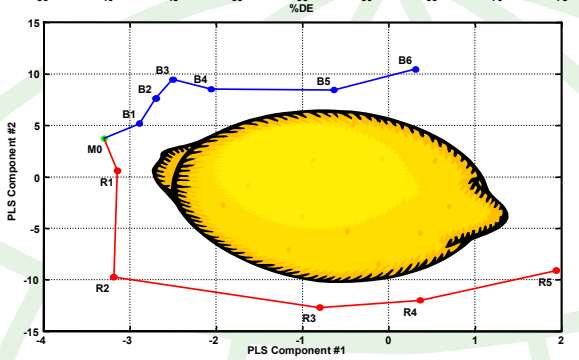
Acknowledgements

The authors would like to gratefully acknowledge the Industrial PhD grant made available by the Ministry of Science, Technology and Innovation. Laboratory technicians Anette Møller, Britta Pedersen and Hanne Thulstrup have made a great effort in developing the method and running the samples.









References

[1] R. Bro; Multi-way calibration. Multi-linear PLS; J. Chemom., 10 (1), 47-62, 1996.

[2] C. A. Andersson and R. Bro; The N-way Toolbox for MATLAB, version 2.11; Chemometrics and Intelligent Laboratory Systems. 52 (1), 1-4, 2000; http://www.models.kvl.dk/

[3] R. Bro, H. A. L. Kiers and C. A. Andersson; PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts; J. Chemom., 13 (1), 295-309, 1999.

Paper I

Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amidated pectin production

Christian B. Zachariassen, Jan Larsen, Frans van den Berg,
Søren Balling Engelsen

Chemometrics and Intelligent Laboratory Systems, **76** (2005), 149–161

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 76 (2005) 149–161

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amidated pectin production

Christian B. Zachariassen[a,*], Jan Larsen[a], Frans van den Berg[b], Søren Balling Engelsen[b]

[a]*Quality Control Development Group, CP Kelco ApS, Ved Banen 16, DK-4623 Lille Skensved, Denmark*
[b]*Department of Dairy and Food Science, Centre for Advanced Food Studies, Food Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

Received 18 March 2004; received in revised form 13 October 2004; accepted 13 October 2004
Available online 20 December 2004

Abstract

The development of an on-line process monitoring and control system for the automatic dosing of ammonia for production of Low Methoxylated Amidated (LMA) pectin is described. The system is based on on-line optical fibre Near Infra-Red (NIR) transmittance spectroscopy of the amidation liquid, consisting of ammonia, isopropylic alcohol, water and minor impurities. The focus in this paper is on the chemometric calibration development and validation, and the improvement on the ammonia dosing for the production. Feasibility studies revealed that the calibration for ammonia, based on NIR transmittance spectra (spectral region from 5300 to 11 000 cm^{-1}) is highly temperature dependent in the range −9–20 °C, but careful choice of spectral pre-processing and wavelength selection could eliminate the temperature dependency. Using the first derivative of the NIR spectra and the selection of the spectral region 5423–6658 cm^{-1}, the average Partial Least Squares regression prediction error was reduced to 1.75% of the full range of the initial calibration, as determined from an independently validated data set one full year after initial implementation in the factory.
© 2004 Elsevier B.V. All rights reserved.

Keywords: NIR spectroscopy; Ammonia; On-line; Process monitoring and control; Pectin

1. Introduction

Low Methoxylated Amidated pectin (LMA pectin) is typically produced by amidating conventional High Methoxylated pectin (HM pectin) by ammonia in an alcoholic suspension [1–3]. Fluctuations in the feed of ammonia to the amidation reactors lead to less controlled amidation and thereby undesired product quality variation. In the traditional off-line measurements approach, the total consumption of ammonia could not be estimated accurately before amidation took place. Traditionally, the concentration of ammonia in the amidation reactors at the CP Kelco pectin manufacturing site in Denmark (Lille Skensved) was measured once every hour, through titration by skilled operators, in a non-laboratory environment. If judged necessary, the titrations would be followed up by manual dosing of ammonia. For these reasons, it was not feasible to closely monitor and act upon the fluctuations in ammonia concentration with sufficient accuracy. It was anticipated that an automated on-line, higher frequency determination of the ammonia concentration during amidation, coupled with an automated dosing of ammonia would result in a better control of the process. This would in turn improve process capability, product consistency and improve adherence to product specifications. After thorough feasibility studies, it was decided to pursue an on-line Near Infra-Red (NIR) transmittance setup to monitor the amidation reaction as input to a high frequency dosing-control of ammonia in the reactor fluids, i.e. optimize the reaction conditions. Based on the process and factory layout, it was decided to establish on-line NIR measurement points at three different tanks. Tank 1 and Tank 3 are auxiliary tanks; Tank 2 contains the liquids during the amidation reaction. Ammonia can be fed to the process liquids in Tank 2 from external

* Corresponding author. Tel.: +45 5616 5616; fax: +45 5616 9446.
E-mail address: Christian.Zachariassen@cpkelco.com (C.B. Zachariassen).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.10.005

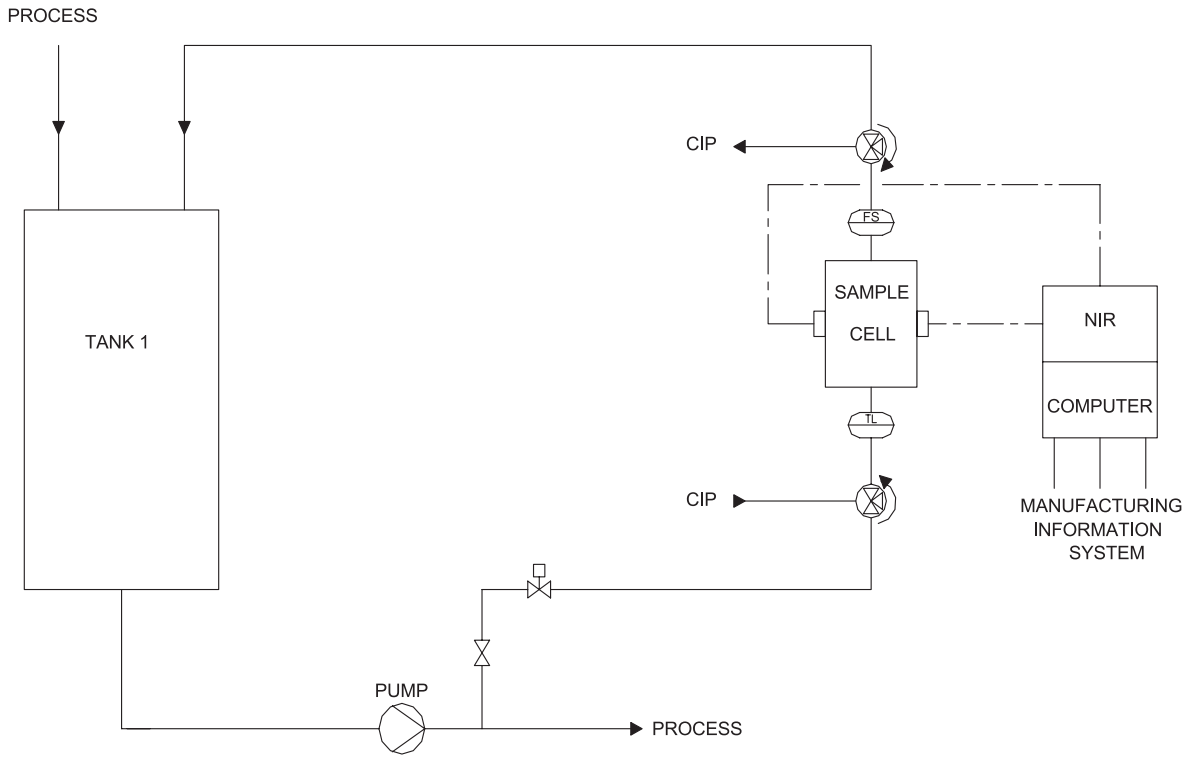


Fig. 1. Layout of the side stream sampling loop for Tank 1. TL denotes a temperature logger and FS denotes a flow switch. The loop can be rinsed by a CIP stream (Cleaning In Place) when valves turn as indicated by arrows.

sources, making it the point most suitable to control the concentration of ammonia. NIR has previously been demonstrated to be useful for rapid off-line quality control of the final LMA pectin [4,5], and by using optical fibres one NIR instrument with multiplexer may control several measuring points. The overall setup pursued was therefore to continuously measure the concentration of ammonia in the three tanks and to supply the measured ammonia concentration to the manufacturing information system, which can be programmed to dose ammonia into Tank 2 by regulating valves.

For process design reasons, it was chosen to measure the amidation liquid in a side stream sampling loop, rather than in the production line or directly in the tank itself. The layout of the side stream sampling loop for one of the measurement points can be seen in Fig. 1.

This paper will focus on the feasibility studies (the "off-line" development) before the installation of the sampling loop, the chemometric calibration and validation of the acquired NIR spectra to the ammonia concentration (the on-line development) and the implementation of the automated ammonia dosing system based on the subsequent on-line determination of ammonia.

2. Materials and methods

Samples are measured with a Bomem NetworkIR FT-NIR spectrometer (a similar instrument is now manufactured under the name FTBA2000-200 by ABB Analytical). The instrument is fitted with 2.7-μm InGaAs detectors. The liquid flows in a PFA Teflon® 12 mm OD (outer diameter)×10 mm ID (inner diameter) tube fixed in a Bomem Side Stream Sample Cell. The sample cell is equipped with a fitting to allow optical path-length adjustment, and with connectors for optical fibre cables. NIR spectra were acquired using BGrams32 v. 4.04 software from Thermo Galactic and CAAP (Continuous Automated Analysis Program) v. 2.1 and 3.0 from Bomem. Reference values for ammonia were determined by titration in triplicate using an internal CP Kelco control method. The multivariate NIR data were analyzed using Guideline 7.6 SR-1 from Camo, PLSplus/IQ 5.09 from Thermo Galactic and Matlab 6.5 (R13) from The MathWorks using the PLS-Toolbox, Version 3.0.3a from Eigenvector Research and the Command Line iPLS toolbox version 2.0 [6].

3. Off-line development

3.1. Experimental

To develop an initial calibration model between NIR spectra and reference values, it was decided to use synthetic samples of amidation liquids prepared in the laboratory, closely resembling process liquids. A ternary mixture design with water, isopropylic alcohol (IPA) and ammonia was

constructed to span the calibration in the range for established normal process operation. Known minor concentrations of impurities (<3% by weight) present in normal process samples were also added. The concentration of the compounds was estimated by weight with a precision error less than 0.01%.

As a calibration set, 19 samples were prepared. A further 11 samples were prepared as an independent test set. The selected variation of IPA in the samples was pseudorandom to reflect the assumed composition of typical process liquids in the three tanks, whereas the variation of ammonia in the calibration set was systematically varied to cover the desired calibration interval. Fig. 2 shows the mixture design. For reasons of confidentiality, the concentrations of IPA and ammonia have been rescaled to 0–100%, which has the inherent advantage that Root Mean Squared Errors will be expressed in dimensionless % of the calibration range, which enables calibration performance comparison to other applications.

Apart from the impurities added, the system can be regarded as a ternary mixture consisting of ammonia, IPA and water. In the laboratory, the same NIR instrument later to be used in the process, as well as the process interface and fibre optic cables were installed. The liquids were cooled to approximately −18 °C and placed in the Teflon® tube mounted in a Side Stream Sample Cell. The liquids were measured in a temperature range starting at the sub-zero level between −9 and −1 °C, and next at 0, 5, 10, 15 and 20 °C. In all, 152 spectra were acquired, as not all samples were measured at all temperatures (primarily sub-zero and 0 °C measurements missing).

3.2. Results and discussion

Absorbance spectra were acquired in the spectral region from 5300 to 11 000 cm^{-1}, corresponding to 1887–909 nm. Below 5300 cm^{-1}, water absorbance was too high for the chosen path length. Above wave number 11 000 cm^{-1}, the signal to noise ratio was too low (the average synthetic spectra can be seen in Fig. 5). To investigate the quantitative performance of the instrumental setup, a Partial Least Squares (PLS) regression model [7,8] of the mean centred spectra in the 5300–11 000-cm^{-1} region with ammonia as response variable was calculated. The score plot of PLS component 1 vs. 2 is shown in Fig. 3.

The score plot clearly reflects the synthetic sample domain from Fig. 2. However, the most striking feature revealed by Fig. 3 is the impact of temperature on the score values. Temperatures within the sample run from low (upper right corner) to high (lower left corner) in the score plot in a systematic way. The temperature effect is often more prominent than the sample-to-sample difference, and if not corrected for, the temperature dependency will deteriorate the calibration performance. In the score plot, high ammonia concentration samples (17–19 and 29) are in the top right corner and low ammonia concentration samples are in the lower left corner, opposing the observed direction of temperature. The effect of increasing temperature and increasing ammonia concentration will therefore cancel out in the score plot. As a consequence, low temperature samples will be predicted to have a higher ammonia concentration, as reflected on

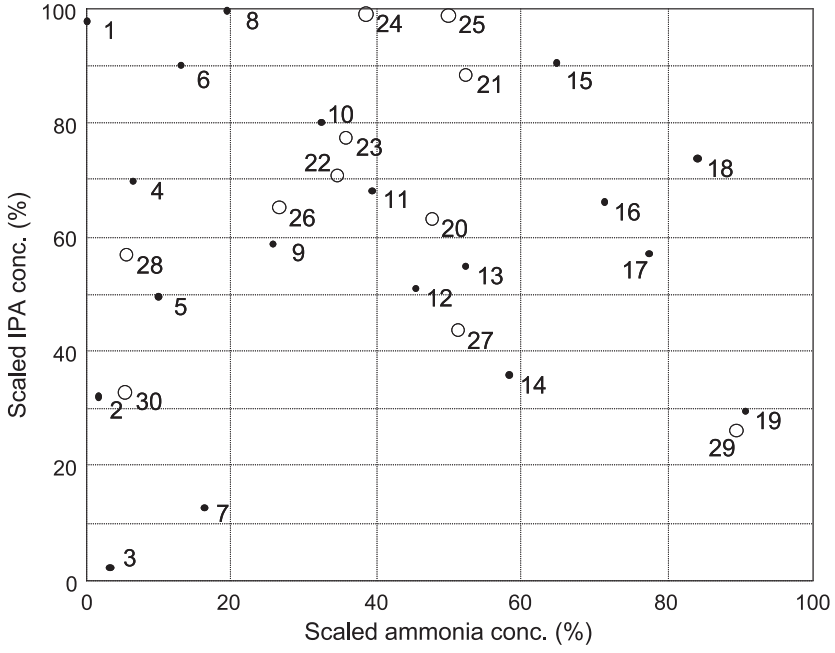


Fig. 2. Synthetic sample domain in two dimensions. ●, Calibration samples; ○, test set samples.

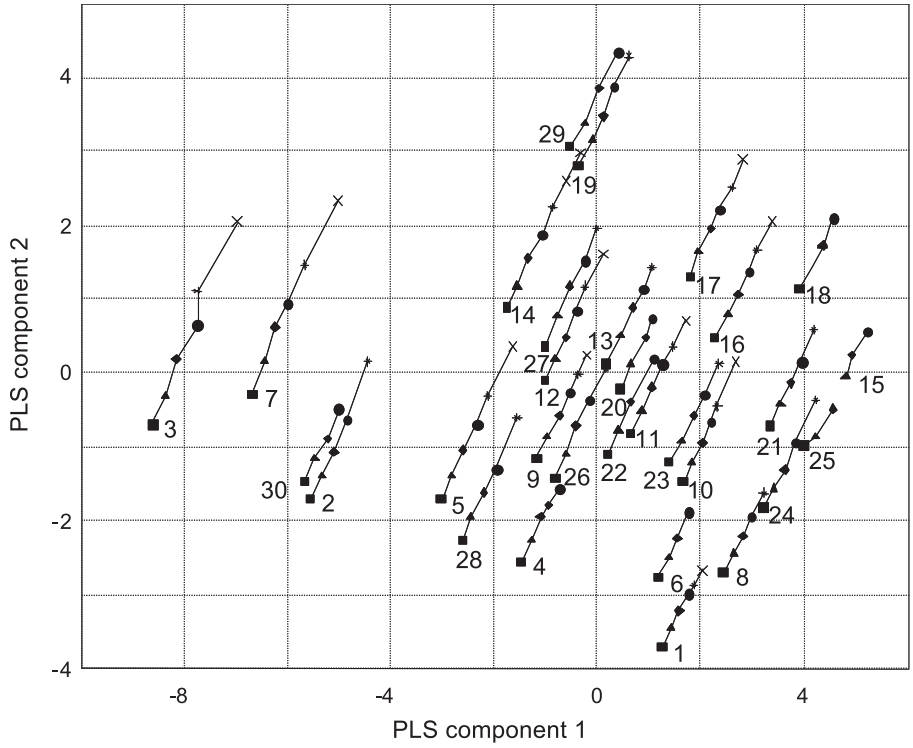


Fig. 3. Score plot of PLS component 1 (75%×-variance explained) vs. 2 (21%) of mean centred spectra (530011 000 cm^{-1}) of the synthetic samples, measured at different temperatures. Numbers indicate samples corresponding to Fig. 2. ×, <0 °C; *, 0 °C; ●, 5 °C; ◆, 10 °C; ▲, 15 °C; ■, 20 °C.

the predicted vs. measured ammonia concentration plot shown in Fig. 4.

Using two PLS components, the Root Mean Squared Error of Calibration (RMSEC) is 9.2%, and the Root Mean Squared Error of Cross Validation (RMSECV) was calculated to be 9.6% using leave-one-sample using leave-one-sample-out (at all temperatures) cross validation. Since the sample set from a chemical point of

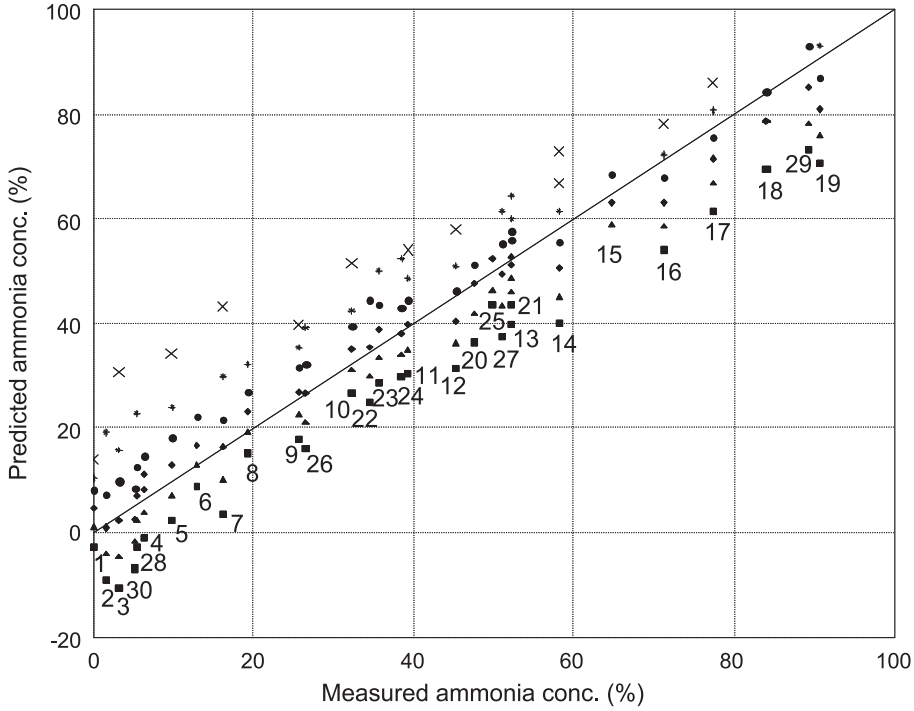


Fig. 4. Predicted (two PLS components) vs. measured ammonia from PLS model on mean centred spectra (5300–11 000 cm^{-1}) of the synthetic samples. A two-component model clearly does not capture the variation in temperature. ×, <0 °C; *, 0 °C; ●, 5 °C; ◆, 10 °C; ▲, 15 °C; ■, 20 °C.

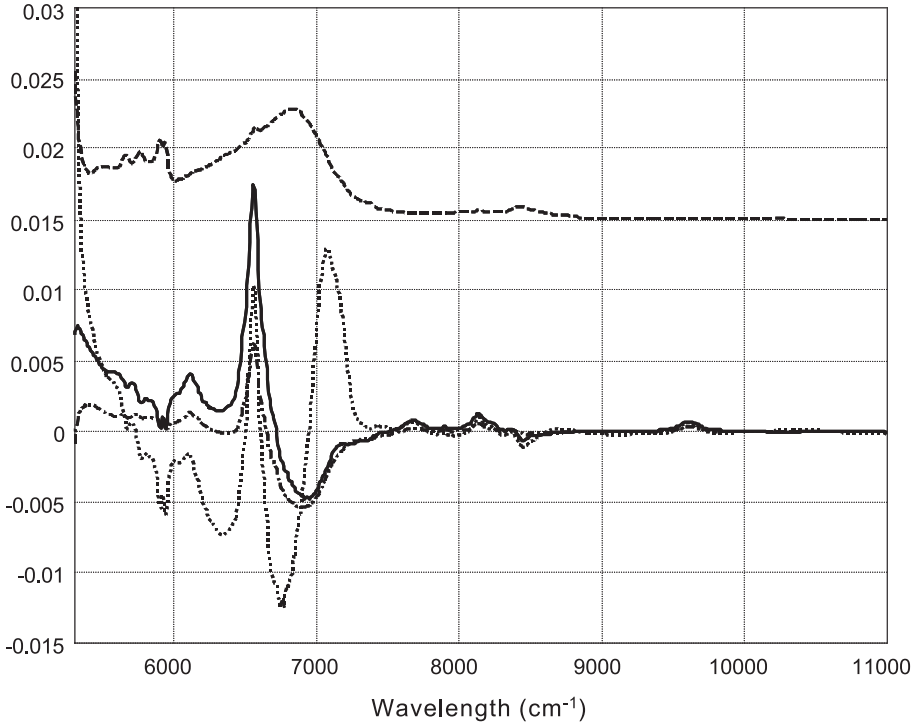


Fig. 5. Loading weights from PLS model on mean centred spectra (5300–11 000 cm^{-1}) of the synthetic samples. The legend reflects a spectroscopic interpretation of the loading weights; scaled mean spectrum (dashed line) shown for interpretation purposes. Component 1 (solid line), Component 2 (dotted line), Component 3 (dash-dot line).

view can be regarded as a three compound mixture, two PLS components should be sufficient to describe this system with closure. However, because the effect of temperature is so drastic, it can be regarded as an independent source of variation. Therefore, including the third PLS component compensates for the effect of temperature. In Fig. 5, the loading weights of the first three PLS components are displayed.

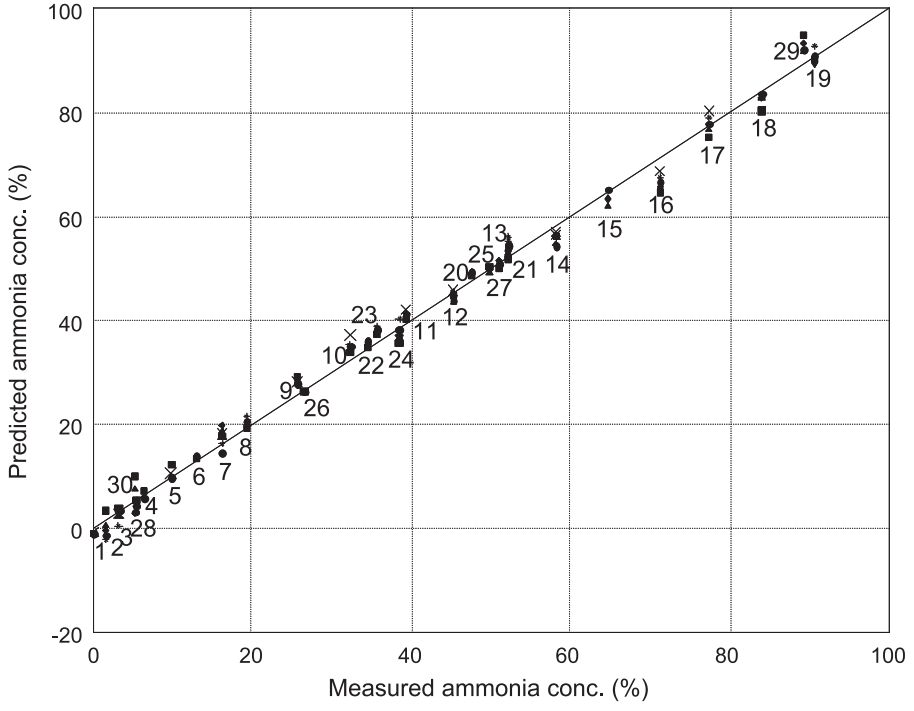


Fig. 6. Predicted (three PLS components) vs. measured ammonia from PLS model on mean centred spectra (5300–11 000 cm^{-1}) of the synthetic samples. A three-component model captures the variation in temperature. ×, <0 °C; *, 0 °C; ●, 5 °C; ◆, 10 °C; ▲, 15 °C; ■, 20 °C.

The loading weights for PLS component 1 are dominated by the ammonia specific peak at 6573 cm^{-1} overlapping the water peak at 6100–7600 cm^{-1}. The loading weights for PLS component 2 have features from the IPA (~5900 cm^{-1}) and the mentioned ammonia and water peaks. The temperature correcting PLS component 3 does not display any distinct features but contains information about the ammonia peak and water peak. It is however clearly connected with the compensation for band-shift as a function of temperature in NIR spectroscopy [9].

Fig. 6 shows the predicted vs. measured ammonia concentration for a model with three PLS components. Compared with the two component predictions shown in Fig. 3, the improvement is obvious. The RMSEC is 2.1%, and RMSECV is reduced approximately four times to 2.3%.

4. On-line development

4.1. Experimental

Direct application of the calibration developed from the synthetic samples to process amidation liquids did not give satisfactory results, even when using the same instrument and the measurement cell used in the laboratory setup. Apparently the synthetic samples did not match and span the variation of the impurities found in the process samples. As an additional complication, it proved impossible to reproduce the exact same path length when mounting the Teflon® tubes in the Side Stream Sampling Cell used in the process environment. For this reason, an extended calibration, based on actual process amidation liquid, was computed in order to take the observed impurities and the different path lengths into account (see Table 1).

As listed in Table 1, substantially more process samples were acquired than the original 30 synthetic samples. Process samples were sampled during routine on-line measurements. As the NIR instrument was scanning through the Teflon® tube, a sample grab was made from a valve directly below the Teflon® tube after a first dead volume was washed out. The sample, approximating the liquid scanned through the Teflon® tube, was subsequently titrated in triplicate for determination of IPA and ammonia concentration by laboratory personnel. Just before the end of the process test set acquisition period the Side Stream Sampling Cell measuring Tank 2 was changed after discovering the optical fibres were not perfectly aligned across from each other. The samples acquired with the new Teflon® cell have been labelled "Tank 2 (new)". Although the design of the sample cell will allow for replacement of the cell without changing the path length significantly, some deviation is anticipated. The method should be robust to such changes.

Table 1
Overview of synthetic samples and process samples acquired, including the expected variation of concentration of ammonia and IPA

	Sampling point	Number of samples	Ammonia concentration	IPA concentrations
Synthetic	Laboratory	30	Full range	Full range
Process calibration set	Tank 1	83	Low	Full range
	Tank 2	83	Intermediate	Low
	Tank 2 (new)	0	Intermediate	Low
	Tank 3	85	High	Low
Process test set	Tank 1	89	Low	Full range
	Tank 2	85	Intermediate	Low
	Tank 2 (new)	3	Intermediate	Low
	Tank 3	89	High	Low

4.2. Results and discussion

The average spectrum for synthetic and process samples for each sampling point mentioned in Table 1 is plotted in Fig. 7; spectra from both the process calibration set and the process test set have been averaged.

The absorbance of Tank 1 samples acquired is higher than the absorbance of samples from other tanks and the absorbance of the synthetic samples measured in the laboratory (this cannot be seen in Fig. 7 however, as the spectra have been offset for clarity). This indicates that the path length of sampling cell 1 is longer than the other. Fig. 7 also shows that the new sampling cell measuring amidation liquid in Tank 2 measures slightly different spectra than the initial cell. The difference in absorbance characteristics is reflected in the PLS score plot of PLS component 1 and 2 shown in Fig. 8.

The score plot separates Tank 1 samples from the rest of the samples in the main variation direction (the first component). The three new measurements from Tank 2 (×) are also separated from the other Tank 2 measurements. Furthermore, the score plot reveals that the different temperatures of the synthetic samples are not handled correctly. A robust model must be able to handle differences in path length as well as changes in temperature. The optimal number of PLS components to predict the ammonia content in the process test set from the synthetic samples and the process calibration set was determined to be four with an RMSEC of 3.6%, and the Root Mean Squared Error of Prediction (RMSEP) was calculated to be 3.3%. The RMSEP is lower than the RMSEC due to the fact that the synthetic samples are included in the calibration set, and the synthetic samples and the temperature are not handled well in the resulting model. Despite the use of four PLS components compared to three for the synthetic sample-set alone, the same level of predictive performance is not reached. Differences in path length and temperature can be accounted for by various pre-processing methods and spectra interval selection [10]. The choice in pre-processing methods is however, in practice, limited by those supported by

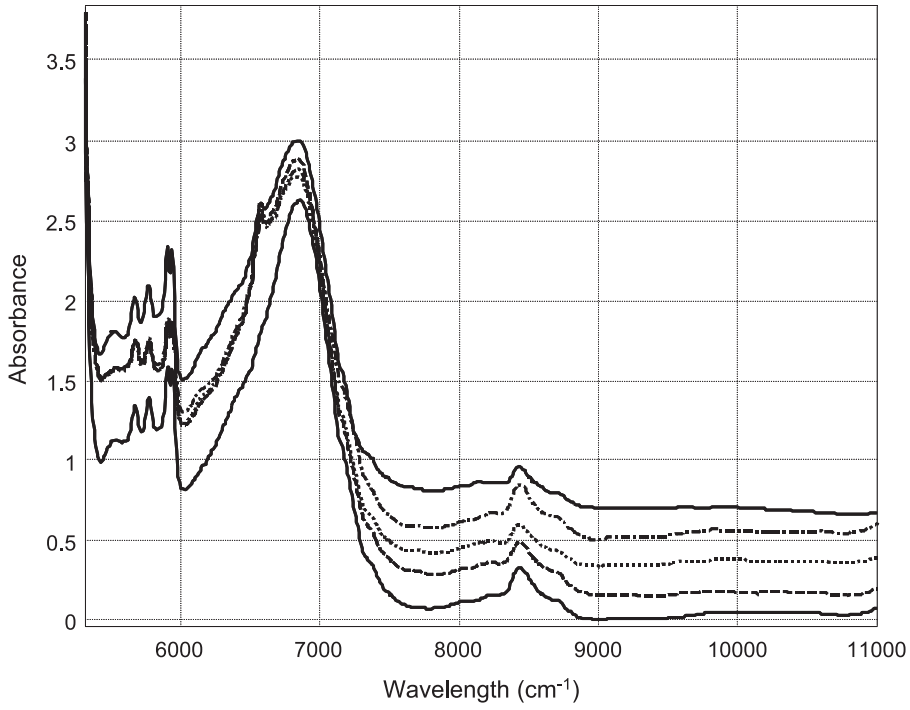


Fig. 7. Average spectrum synthetic and process samples. Synthetic samples (lower solid line), Tank 1 (upper solid line), Tank 2 (dash-dotted line), Tank 2 new (dashed line), Tank 3 (dotted line). Spectra have been offset on the absorbance scale for clarity.

CAAP—the Continuous Automated Analysis Program—the software to be used during process operation. This restricts the pre-processing methods to direct arithmetic on the acquired spectra (e.g. baseline subtraction, area normalization, mean centering and variance scaling), smoothing and derivatives by Stavitsky–Golay algorithm [11]. For modelling purposes, variable selection, pre-processing, outlier detection and calibration are all

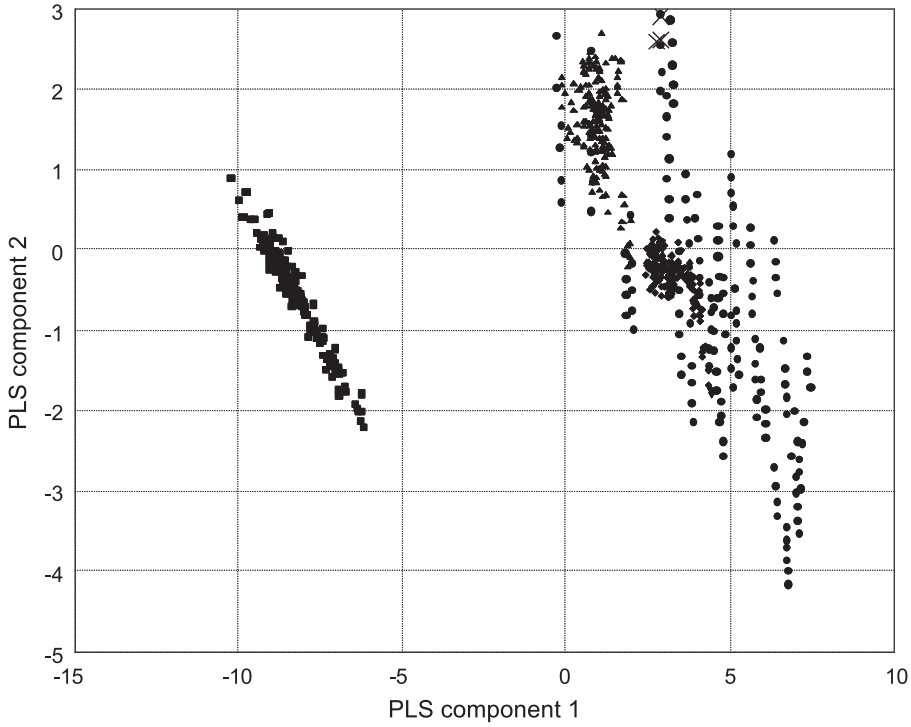


Fig. 8. Score plot of PLS component 1 (88%×-variance explained) vs. 2 (6%) of mean centred spectra (5300–11 000 cm^{-1}) of synthetic and process samples collected at different locations. ●, Synthetic samples; ■, Tank 1; ◆, Tank 2; ×, Tank 2 (new); ▲, Tank 3.

interdependent phenomena. Therefore, an iterative procedure, where all aspects are continually revised, is recommended [12]. One approach to spectral region selection is interval PLS (iPLS) [13]. The entire spectral region measured is divided into a predefined number of evenly spaced intervals, and local PLS models are fitted to each interval using from one to a selected number of PLS components (called local models). In this work, iPLS models were fitted using the mean centred full region (5300–11 000 cm^{-1}) spectra with no further pre-processing and the results of the analysis are given in Fig. 9.

Since the iPLS routine does not support test set validation [6], the local models and the global (full region) model are validated by cross validation with two segments: segment one consists of the synthetic and the process calibration set and segment two consists of the process test set. The RMSECV of the local models is compared to the RMSECV of the four PLS component global model, being the one with the lowest RMSECV (4.4%). Due to the different validation approaches, the RMSECV is not directly comparable to the RMSEC or RMSEP of the PLS model shown in Fig. 8.

From Fig. 9, it is observed that one local model gives an RMSECV lower than that of the global model. It is based on interval number five (counting from left to right in Fig. 9) using three PLS components, spanning the region 6442–6720 cm^{-1}, which gives an RMSECV of 4.0%. From the four PLS components local model plot in Fig. 9, the intervals 4–5 (6156–6720 cm^{-1}), 9 (7584–7861 cm^{-1}), 11 (8154–8432 cm^{-1}) and 15–16 (9296–9859 cm^{-1}) can be identified as being important as they have nearly as low an RMSECV as the global model. The intervals include spectral regions, which may be assigned to N–H stretch first overtone, N–H combination bands and N–H stretch second overtone [14]. All these regions are important for aqueous solution, where water is available for hydrogen bonding. These results from iPLS may be iteratively combined with spectroscopic and chemometric knowledge to identify combinations of pre-processing, spectral region selection, outlier detection and choice of number of PLS-components to yield consistent, low prediction errors. Rather than giving a final evaluation of expected prediction error to a chosen model based solely on the calibration set, the process test set is used

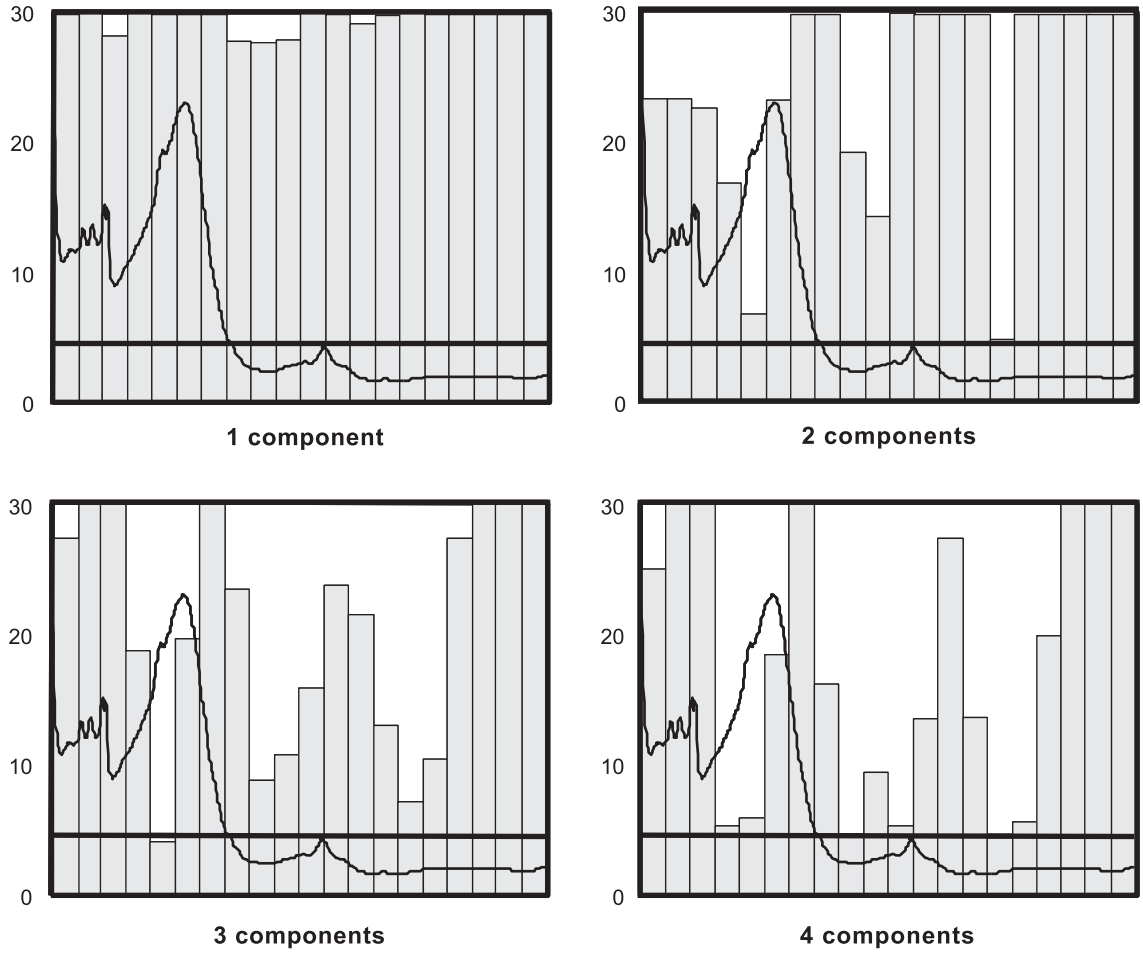


Fig. 9. iPLS plots with the bars showing RMSECV of local models using one to four components. Horizontal line indicates RMSECV for a full spectrum, four-component model and the background spectrum is the average of all synthetic and process spectra.

Table 2
Performance of selected models

Pre-processing	Region (cm^{-1})	Outliers	Number of PLS-components	RMSEC (%)	RMSEP (%)
None	5300–11 000	0	4	3.55	3.27
None	5300–11 000	3	4	3.55	2.62
None	5423–9011	0	4	4.07	2.90
None	6442–6720	0	3	2.60	2.83
None	7252–10 770	0	5	1.95	3.60
MSC	5423–9011	0	4	3.55	3.05
MSC	7252–10 199	0	4	3.69	3.88
MSC	5423–6658	0	2	5.03	4.26
MSC	5423–6025	0	5	4.70	5.09
1Dev SG 552	5423–10 199	0	4	2.01	1.87
1Dev SG 552	5423–10 199	8	4	1.68	1.79
1Dev SG 552	7252–10 199	0	3	2.70	2.23
1Dev SG 552	7252–10 199	2	3	2.53	2.20
1Dev SG 552	6442–6720	0	3	2.17	2.27
1Dev SG 552	5423–6658	0	3	1.98	1.90
1Dev SG 552	**5423–6658**	**6**	**2**	**1.66**	**1.86**
2Dev SG 552	5423–10 199	5	3	1.76	1.94
2Dev SG 552	7252–10 199	6	4	2.53	2.64
2Dev SG 552	6442–6720	6	3	2.18	2.45
2Dev SG 552	5423–6658	5	2	2.56	2.46

The model highlighted in bold is the chosen model. MSC refers to "Multiplicative Signal Correction" [15], 1Dev SG 552 and 2Dev SG 552 refer to first and second derivative computed by Savitzky–Golay algorithm [11] using five points left and right and fitting a second-degree polynomial.

for this purpose. The results in terms of prediction errors of some of the models are listed in Table 2. The choice of pre-processing methods is quite restricted, since only a few methods were supported in the on-line CAAP software.

As indicated in the table, the greatest impact on prediction error comes from spectral region selection. First derivative pre-processing proved better than other pre-processing methods, other choices being equal. The first derivative spectra displayed in Fig. 10 clearly—in this case—have removed the baseline offset of the raw spectra, and thus give good compensation of differences in path length. Other methods for removing baseline offset (e.g. normalization) could also be considered.

The region 5423–6658 cm^{-1} gives the best overall model, and includes the first overtone of both IPA and ammonia. The 5423–6658 cm^{-1} region has absorbance values of 0.8 to 2.8, beyond what is usually normally accepted as the linear range of calibrations. The region also ends at the shoulder of the water peak at 6100–7600 cm^{-1}, which could be questionable, but from Table 2 it is seen that inclusion of the complete water band at 6100–7600 cm^{-1} does not improve the performance of the calibration models. A shorter path length could be considered, but would require custom-made cell housing and different choice of Teflon tubing. The score plot from a model based on the first derivative of the spectra in the region 5423–6658 cm^{-1} can be seen in Fig. 11.

The synthetic sample domain from Fig. 2 can readily be recognized, albeit rotated slightly counter-clockwise. The PLS component one can be interpreted as ammonia concentration and PLS component two can be interpreted as the IPA concentration, i.e. the position of the process samples in the score plot thus reflects their IPA and ammonia content. The process engineers actually over-estimated the IPA concentrations and fluctuations in Tanks 2 and 3 before conducting this study. Therefore, the synthetic samples proved not to cover the entire variation exhibited by the process samples. In contrast, they supplement the variation, which was beneficial for the calibration. As seen in Table 2, the calibration has RMSEC of 1.66% and RMSEP of 1.86% using two PLS components. The RMSEP actually has a minimum of 1.82% using three PLS components, but the more parsimonious (and therefore presumably more robust) two-component model was chosen in favour of an estimated gain of 0.04% in predictive ability.

5. Implementing on-line control

For the on-line implementation, a PLS model equivalent to the before mentioned two component model based on the region 5423–6658 cm^{-1} was computed using the combined calibration and test set. The combined model had an RMSEC of 1.75%. When the on-line NIR ammonia predictions were compared to the manual titrations done by process operators, the results were surprising. The process operators routinely analyzed a sample approximately once an hour to monitor the ammonia concentration. If the level was assessed as low or if a new amidation was forthcoming, the operators would feed ammonia to main Tank 2 by manual operation, based on process experience. The operational effect would be evaluated from the next titration 1 h later, as titrating a sample immediately after dosing would give an "unreliable" (i.e. out of specifications) result. Fig. 12 shows the results of the NIR predicted concentrations superimposed on the process titrations. When only the process titrations are considered, the process appears stable and within the specified ammonia concentration set point. The NIR predicted results however reveal that the dosing of ammonia by the operators results in strong concentration increase (reaching levels of more than 15% above target point), which are rapidly consumed in the amidation reaction. The magnitude of the concentration peaks and the rapid consumption were unknown to process

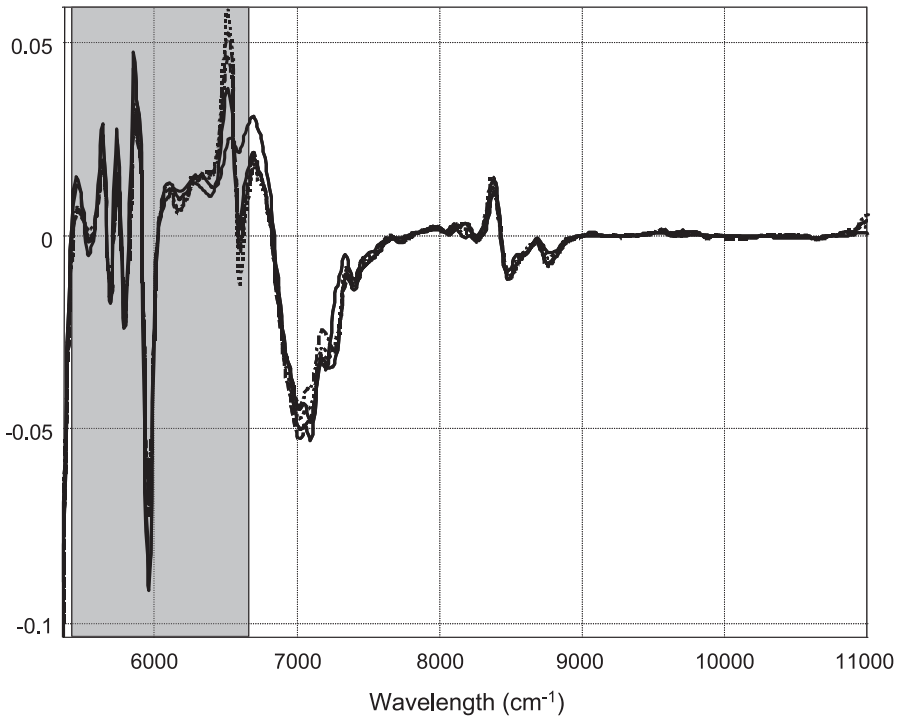


Fig. 10. Average first derivative of synthetic and process spectra. The gray area marks the chosen wave number region 5423–6658 cm^{-1}. Synthetic samples (solid line), Tank 1 (solid line), Tank 2 (dash-dotted line), Tank 2 new (dashed line), Tank 3 (dotted line).

engineers before implementing the process analytical measurement system.

Fig. 12 reveals that the NIR predictions are extremely precise and very stable from one measurement to the next (best observed in a time span where no dosing of ammonia occurs). This implies that the spectroscopy plus process interface as well as the PLS model are very robust, and noise levels are very low.

After an evaluation period, the NIR predictions proved reliable and process control engineers phased in an

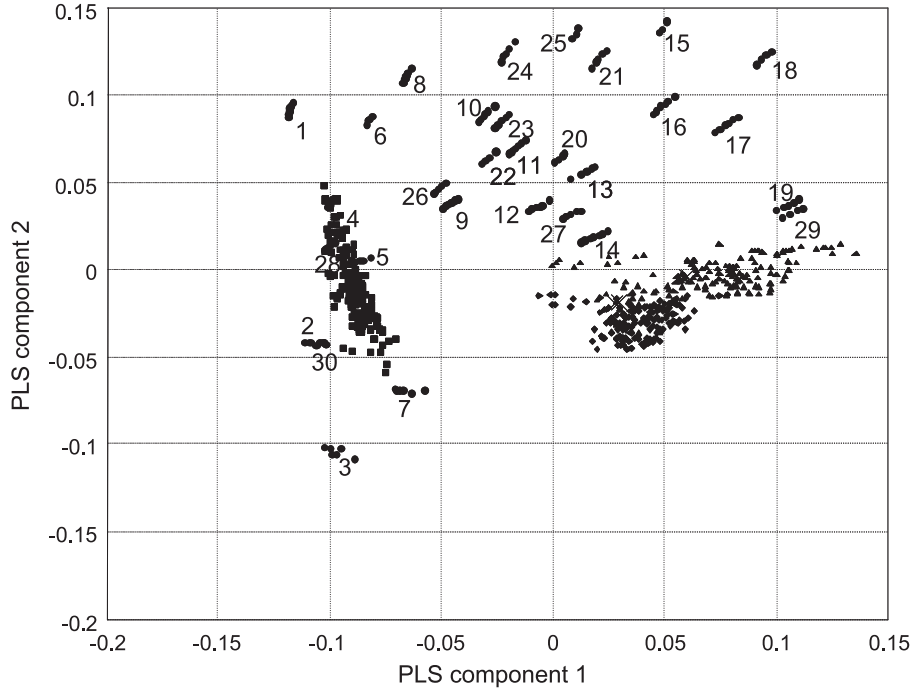


Fig. 11. Score plot of PLS component 1 (71%×-variance explained) vs. 2 (24%) of mean centered first derivative spectra (5423–6658 cm^{-1}) of synthetic and process samples. The synthetic samples are numbered in accordance with Fig. 1. ●, Synthetic samples; ■, Tank 1; ◆, Tank 2; ×, Tank 2 (new); ▲, Tank 3.

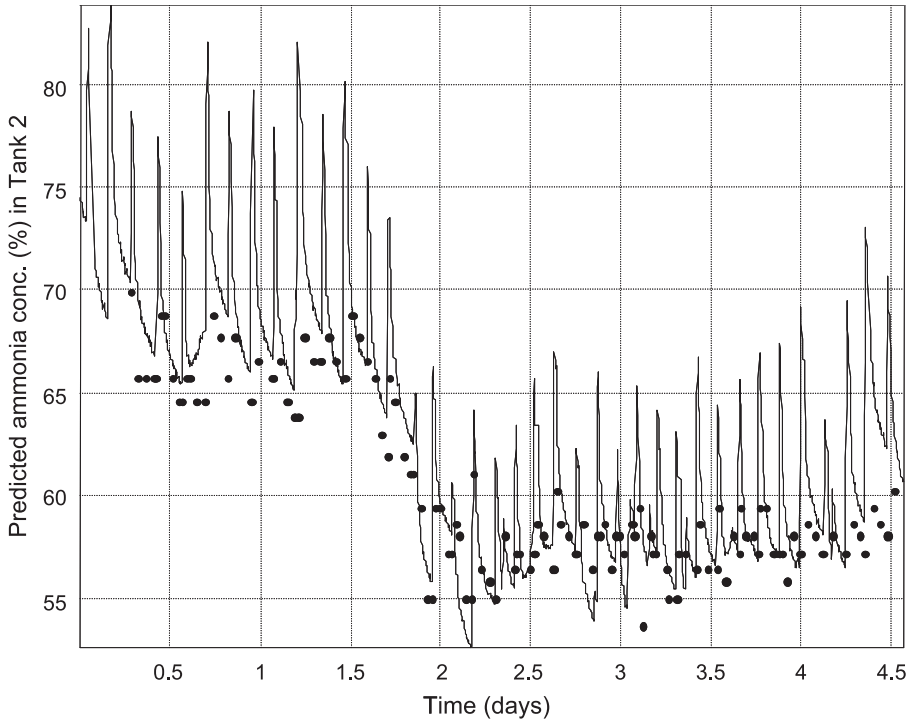


Fig. 12. NIR predicted values superimposed on manual ammonia titrations. The titrations done once an hour do not capture the full process dynamics. At 2 days, the set point for the ammonia concentration changed due to processing conditions. ●, Process titrations; solid line, NIR predicted values.

automated dosing system based on the NIR predictions. An automatic valve, capable of dosing ammonia in short bursts, replaced the manual valve. The automatic dosing was introduced slowly over a period of 14 days in order to gather experience and not to upset production quality. Fig. 13 shows NIR predicted ammonia values from the production during that period. It is clearly observed that the ammonia concentration is much more in control,

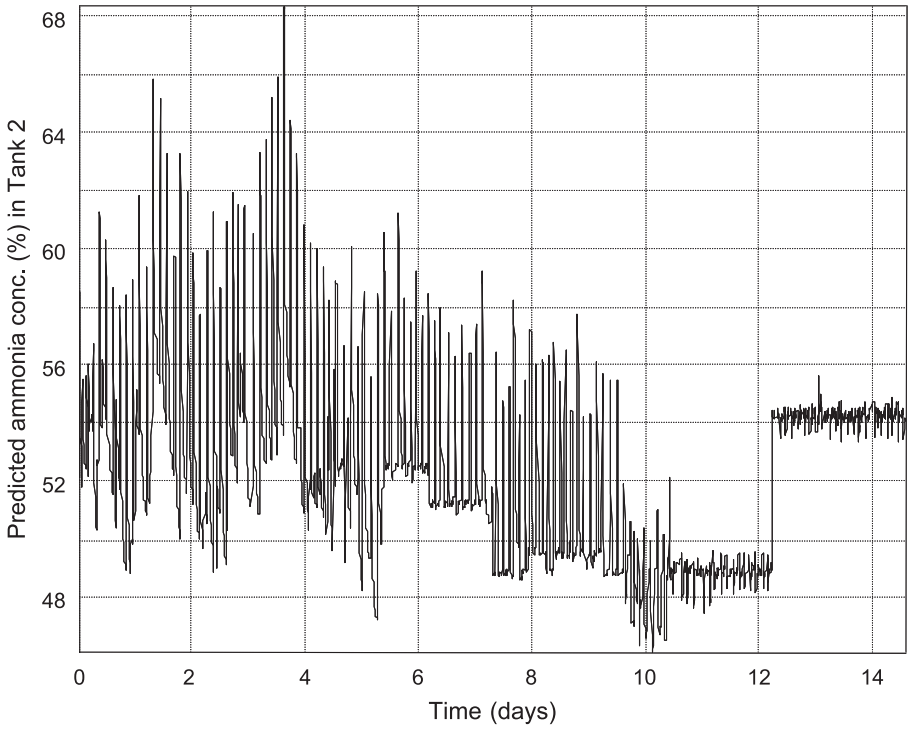


Fig. 13. NIR Predicted ammonia concentrations in Tank 2 for 14 days where on-line control of the ammonia concentration was implemented. At 12 days, the set point for the ammonia concentration was altered.

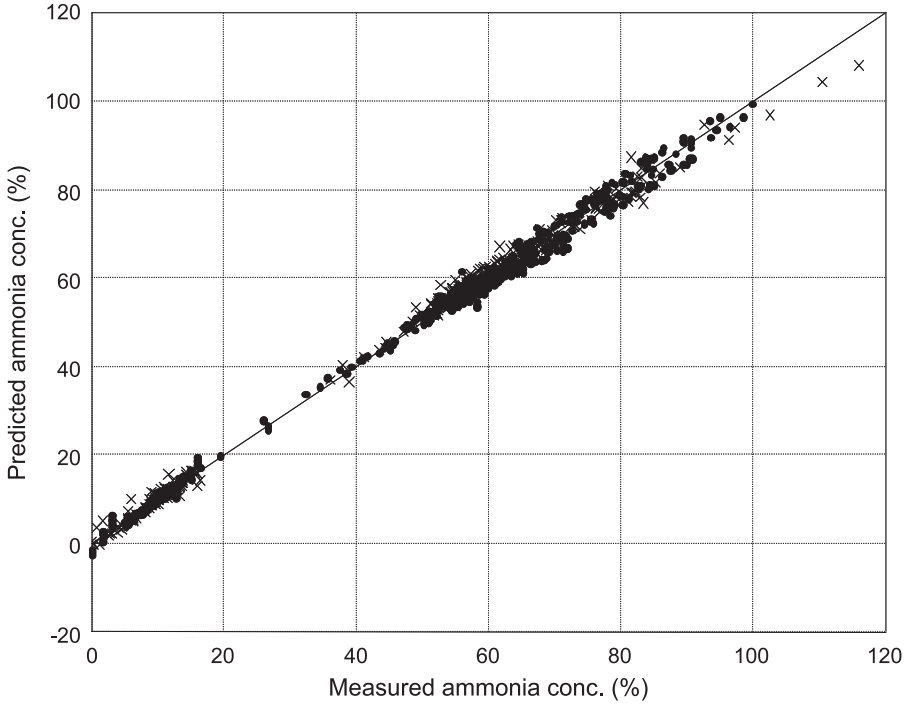


Fig. 14. Predicted (two PLS components) vs. measured ammonia, one full year after implementation of the calibration. ●, Calibration samples; ×, process samples acquired after implementation of the calibration.

being kept at an almost fixed level with near continuous automatic dosing rather than manual dosing every second or third hour.

Concentration set points can now be readily changed within minutes, compared to several hours before introduction of the new system (compare Fig. 13 with Fig. 12).

The long-term performance of the NIR predictions is continuously being evaluated by comparing NIR predicted values to manually retrieved samples titrated by laboratory personnel. The following year, 433 process samples were tested using the implemented PLS model, the results of which can be seen in Fig. 14.

As observed from the figure, the calibration is stable and the new RMSEP is 1.75% with a bias of −0.15%, practically identical to the target RMSEC. Addition of the third PLS component still would give a minimal improvement of RMSEP to 1.74%, but a two-component model is still preferred. Closer inspection of Fig. 14 shows a split in both calibration and validation samples in the range from 60% to 100% ammonia. Approximately half of the samples are predicted to have 2–4% lower ammonia than the rest in that concentration range. The reason for that has not yet been identified. During the evaluation year, it occurred that the ammonia-dosing valve at one stage froze in the open position, with an uncontrolled dosing of ammonia as a result. Samples were retrieved for validation purposes and can be seen in Fig. 13 as validation samples in the range of 90–115%.

6. Conclusion

The implementation of the on-line process monitoring and control of ammonia by NIR spectroscopy at CP Kelco has been very successful. The ammonia prediction performance has been very good, exceeding expectations, and the stability and accuracy of the multivariate calibrations are much better than expected. At all levels of the organization, the predictions are regarded with high confidence. Despite this, process operators on all shifts still acquire samples for titrations—not only to continuously validate the NIR predictions, but also to maintain the ability for the operators to continue production in the event of a breakdown of the ammonia dosing system.

The benefits of the improved control of the ammonia dosing system have been numerous. It is now possible to regulate the concentration of ammonia very precisely and reproducibly. This removes an important source of variation from the amidation reaction as well as a possibility to make minute changes to ammonia concentration to improve product quality. If troubleshooting of the production becomes necessary, the process engineers can assume and quickly verify that the ammonia concentration has been constant.

Implementing this method has also improved control of the ammonia concentration in the amidation liquid, thereby increasing the capability to produce an amidated pectin product with a desired degree of amidation.

All in all, the mentioned benefits have made the application a good investment for CP Kelco.

Acknowledgements

The authors would like to gratefully acknowledge the grant made available by the Ministry of Science, Technology and Innovation to partly cover the expenses in connection with participation in the Industrial PhD project, which is a joint effort between CP Kelco and the Food Technology Group, The Royal Veterinary and Agricultural University of Denmark.

At CP Kelco, many colleagues have contributed in taking this application from idea to reality. The laboratory personnel, viz. Vicky Willumsen, Heidi Olsen, Anette Møller and Karen Christiansen did much of the great effort in sampling and titrating all the samples. A joint effort from construction, automation, IT and process engineers and technicians at CP Kelco made it possible to install the hardware on the production line and provide invaluable advice and suggestions along the way.

All the staff and students at the Food Technology Group have always been available for interesting and thoughtful discussions.

References

[1] E.F. Bryant, Acid amides of pectinic acid and process for their preparation, US Patent, 2,480,710 (1945).
[2] G.H. Joseph, A.H. Kieser, E.F. Bryant, High-polymer, ammonia-demethylated pectinates and their gelation, Food Technology 3 (1949) 85–90.
[3] Compendium of Food Additives Specifications/add. 9, FAO Food and Nutrition Paper no. 52, Rome (2001).
[4] S.B. Engelsen, L. Nørgaard, Comparative vibrational spectroscopy for determination of quality parameters in amidated pectins as evaluated by chemometrics, Carbohydrate Polymers 30 (1996) 9–24.
[5] S.B. Engelsen, E. Mikkelsen, L. Munck, New approaches to rapid spectroscopic evaluation of properties in pectic polymers, Progress in Colloid and Polymer Science 108 (1998) 166–174.
[6] L. Nørgaard, 2000. The command line iPLS toolbox for MATLAB, ver 2.0, KVL, Denmark, http://www.model.kvl.dk/source/ipls.
[7] M. Sjöström, S. Wold, W. Lindberg, J.A. Persson, H. Martens, A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables, Analytica Chimica Acta 150 (1983) 61–70.
[8] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[9] L.G. Thygesen, S.O. Lundqvist, NIR measurement of moisture content in wood under unstable temperature conditions: Part 1. Thermal effects in near infrared spectra of wood, Journal of Near Infrared Spectroscopy 8 (2000) 183–189.
[10] F. Wülfert, W.T. Kok, O.E. de Noord, A.K. Smilde, Linear techniques to correct for temperature-induced spectral variation in multivariate calibration, Chemometrics and Intelligent Laboratory Systems 51 (2000) 189–200.
[11] A. Savitsky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Analytical Chemistry 36 (1964) 1627–1639.
[12] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Elsevier, 1998.
[13] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Applied Spectroscopy 54 (2000) 413–419.
[14] B.G. Osborne, T. Fearn, P.H. Hindle, Practical NIR Spectroscopy with Applications in Food and Beverage Analysis, Longman Scientific and Technical, 1993.
[15] T. Isaksson, T. Næs, The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy, Applied Spectroscopy 42 (1988) 1273–1284.

Paper II

Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system

Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Rasmus Bro, Anna de Juan, Romà Tauler

Chemometrics and Intelligent Laboratory Systems, **83** (2006), 13–25

Available online at www.sciencedirect.com
SCIENCE DIRECT
Chemometrics and Intelligent Laboratory Systems 83 (2006) 13–25
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system

Christian B. Zachariassen [a,b,*], Jan Larsen [a], Frans van den Berg [b], Rasmus Bro [b], Anna de Juan [c], Romà Tauler [d]

[a] Quality Control Development, CP Kelco ApS, Ved Banen 16, DK-4623 Lille Skensved, Denmark
[b] Department of Food Science, Quality and Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[c] Analytical Chemistry Department, Universitat de Barcelona, Martí i Franqués, 1-11, Barcelona, E-08028, Spain
[d] Department of Environmental Chemistry, IIQAB-CSIC, Jordi Girona 18, Barcelona, E-08034, Spain

Received 9 September 2005; received in revised form 24 November 2005; accepted 20 December 2005
Available online 9 March 2006

Abstract

An analytical liquid dilution system is proposed to collect information on the intramolecular distribution of ester groups on the pectin carbohydrate backbone by injecting a solution of pectin into a carrier stream containing a fixed concentration of dye. The dye binds site-specifically to the poly-α-(1→4)-D-galacturonic acid units which are not esterified. The carrier stream is led to a Continuously Stirred Tank Reactor (CSTR). The pectin is slowly diluted while UV–VIS spectra are continuously recorded at the reactor outlet, providing a wavelength-by-time matrix for every sample. Thirty-one pectins, enzymatically deesterified to give different ester distributions, are measured using remethylated lemon pectin as starting material. The combined sample matrices are analysed using multi-sample Multivariate Curve Resolution (MCR) and PARAFAC2, both Alternating Least Squares (ALS) regression algorithms. Both algorithms successfully describe the data giving a lack of fit between 2.0% and 8.3%, depending on the number of components extracted and the choice of constraints applied. The paper discusses how and on which terms the two algorithms can be compared. Interpretable information about the deesterification pattern of the pectin-related species and the extent of deesterification is obtained from the resolved dilution concentration profiles. The most detailed chemical description of the pectin-dye species is obtained with an unconstrained MCR-ALS five-component model.
© 2006 Elsevier B.V. All rights reserved.

Keywords: PARAFAC2; Multivariate curve resolution; Pectin; Blockiness; Intramolecular distribution of esterification; Analytical liquid dilution system

1. Introduction

Pectin is a polysaccharide found in the primary cell walls and intracellular regions of higher order plants. It is a structural element giving both strength and flexibility to plant material. Pectin has excellent gelling, thickening and stabilizing properties and is widely used for commercial applications, primarily in the food industry [1–3]. Pectin mainly consists of a linear chain of poly-α-(1→4)-D-galacturonic acid with varying degrees of esterification (%DE). The distribution of the methyl ester groups on the pectin carbohydrate backbone is an important structural parameter that affects the pectin functionality in customer applications.

However, the exact distribution of ester groups on the pectin backbone is a property that can only be defined for an individual pectin polymer, and this information is as such of no commercial interest, as it is not possible to extract or produce identical pectin polymer molecules from natural sources. On the other hand, a method for measuring average intramolecular distribution of ester groups in bulk pectin is highly desired.

Different pectins extracted from natural lemon peel are investigated. In order to span the sample set beyond natural variation, 31 samples are prepared from a remethylated (i.e. esterified) pectin which is subsequently enzymatically deesterified by two different enzymes, known to deesterify pectin in either a "Random" or "Blocky" fashion. The terms Random and Blocky are loosely defined terms, but it has been suggested to define a blocky group as

* Corresponding author. Quality Control Development, CP Kelco ApS, Ved Banen 16, DK-4623 Lille Skensved, Denmark. Tel.: +45 5616 5616; fax: +45 5616 9446.
E-mail address: christian.zachariassen@cpkelco.com (C.B. Zachariassen).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.12.010

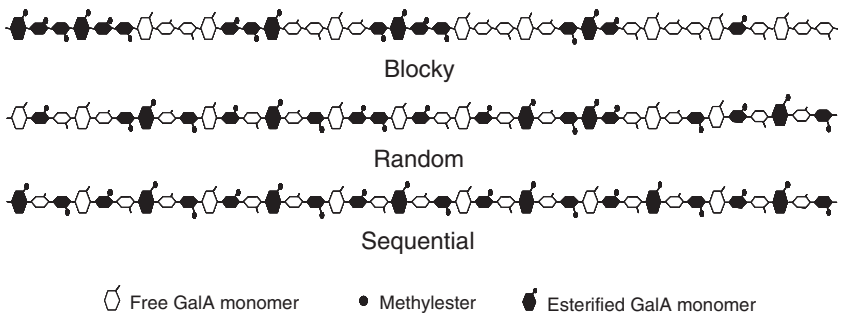


Fig. 1. Idealized diagram of three sections of pectin backbone (GalA = galacturonic acid). The pectins are deesterified 50% but the distribution of estergroups is different (modified from [7]).

a sequence of four or more deesterified galacturonic acids next to each other in the pectin carbohydrate backbone [4–6], while Random groups in this context may cover truly random or more systematic (sequential) deesterification patterns, see Fig. 1.

The extent of randomness or blockiness is a property not easily defined, because the average total degree of esterification (%DE) and the distribution of the ester groups will vary considerably from sample to sample. Additionally the actual degree of esterification and the distribution of the ester groups will vary from molecule to molecule within a pectin sample. For this reason, it is proposed in this article to explore the degree of esterification and the distribution of the ester groups by measuring Ultraviolet–Visible (UV–VIS) spectra. A solution of pectin measured at different time points in a system dynamically diluted with a solution of a fixed concentration of a dye, known to bind site specifically only to the free acid groups of the pectin. This will provide averaged information from all pectin molecules in the solution. A Diode Array Detector (DAD) is used to acquire an entire spectrum for each time point. The spectra recorded will reflect the concentration gradient of the pectin as it is diluted, and every sample will provide a matrix with all the spectra of the pectin sampled. The matrix will have the format wavelength × time and it will contain the absorbance recorded. Such a system has the properties of a Continuously Stirred Tank Reactor (CSTR) and, analytically, will have features known from Flow Injection Analysis (FIA) and conventional High Performance Liquid Chromatography (HPLC), in the sense that the analyte is retained while subjected to chemical reactions. The data can be analysed taking advantage of the features linked to the analysis of multi-way data sets [8,9], which offer the potential to resolve the sought profiles uniquely and to provide quantitative information. In this sense, the species associated with blockwise deesterification can potentially be identified independently from random or sequential deesterification.

This article will present the results of resolving the raw spectra acquired for each sample into meaningful contributions, pure spectra and pure time-dependent dilution profiles, by using the algorithms PARAFAC2 [10,11] and Multivariate Curve Resolution using Alternating Least Squares (MCR-ALS) regression [9,12,13], applying appropriate constraints. The mathematical models will be evaluated with respect to fit and chemical interpretability. The best models identified will be applied to quantify the degree and distribution of ester groups in a second, related paper [14] where also the concepts of random and blocky deesterification are further elaborated on. The same sample set has also been used in a quantitative study of the degree of blockiness using ^{1}H NMR spectroscopy [7].

2. Experimental

2.1. Sample preparation

The samples investigated are laboratory prepared materials derived from a natural lemon pectin — named the "Mother pectin" — produced at CP Kelco, Denmark with a %DE = 72.3 and an average molecular weight of 180 kDa, determined by size exclusion chromatography (SEC). The pectin "Remethylated" was remethylated to %DE = 93.8 by taking 1027 g mother pectin and suspending it in 2000 mL methanol and 150 mL thionyl chloride at 0–5 °C for 5 days. The %DE was verified by titration using an internal CP Kelco control method. The average molecular weight was determined to be 13 kDa using SEC. This corresponds to a degree of polymerization of approximately 50 galacturonic acid units. The remethylated pectin is subsequently deesterified once or twice to give samples numbers 1–31 according to Table 1.

Identification in Table 1 is based on the following convention: from the remethylated pectin (Sample number 33) the samples 1, 5, 9, 12 and 14 are produced by deesterification with a random methyl esterase Rapidase FP Super® from DSM, The Netherlands, known to deesterify in a predominantly random fashion. The samples are given the names "R" (for Random) followed by the degree of *de*esterification (%DDE) induced — not to be confused with the degree of esterification (%DE) defined previously and commonly used to describe pectin. All samples (1, 5, 9, 12 and 14) are produced directly from the remethylated pectin. The five samples produced are then used as starting material for a deesterification process with a block methyl esterase derived from papaya fruit, known to deesterify in a predominantly blockwise fashion. In this step samples 2–4, 6–8, 10–11, 13 and 15 are produced. The samples are named from the starting material followed by the degree of blockwise (B) deesterification. For example, sample number 10 with the name "R21.9 B9.7" is produced by deesterifying the remethylated pectin 21.9% in a random fashion (producing the sample "R21.9 B0.0"), followed by deesterifying that sample 9.7% in a blockwise fashion. The resulting degree of esterification is therefore the %DE of the remethylated pectin minus the degree of deesterification induced,

Table 1
Overview of the sample-set used in the present study

33: Remethylated pectin	32: Mother pectin			
1: R4.9 B0.0	5: R13.8 B0.0	9: R21.9 B0.0	12: R28.9 B0.0	14: R35.0 B0.0
2: R4.9 B10.6	6: R13.8 B7.8	10: R21.9 B9.7	13: R28.9 B13.2	15: R35.0 B6.8
3: R4.9 B18.0	7: R13.8 B13.8	11: R21.9 B21.6		
4: R4.9 B24.5	8: R13.8 B23.8			
16: B4.4 R0.0	21: B14.1 R0.0	25: B22.6 R0.0	28: B33.5 R0.0	30: B36.8 R0.0
17: B4.4 R8.3	22: B14.1 R10.2	26: B22.6 R11.0	29: B33.5 R9.5	31: B36.8 R6.4
18: B4.4 R15.6	23: B14.1 R20.3	27: B22.6 R23.9		
19: B4.4 R22.8	24: B14.1 R30.6			
20: B4.4 R31.9				

The sample number is followed by the sample name. The sample names are interpreted in the following manner: Sample number 19 "B4.4 R22.8" is first deesterified 4.4% by blockwise deesterification and subsequently 22.8% by random deesterification. It is produced by random deesterification of sample number 16, which, in turn, is produced from the remethylated pectin.

in this example 93.8%−21.9%−9.7%=62.2%. The amount of deesterification induced is verified by titration for all the samples. The standard deviation of the %DE titration method is 0.7% in absolute values. The samples 16–31 are produced in a likewise manner, the only difference being that they are initially subjected to a blockwise deesterification followed by a random deesterification. This way all samples produced are subjected to a maximum of two deesterification steps. The range of %DE investigated is from 47.3% to 93.8%.

2.2. Instrumental setup

A 0.5% (weight by volume) solution of pectin is injected by a sample loop connected to an injection valve into a carrier stream flowing continuously and containing a dye in a fixed concentration. The stream leads into a continuously stirred tank reactor with a fixed volume, large compared to the flow rate of the carrier stream. The reactor is initially filled with the carrier stream containing the dye. The injection volume of the pectin solution is small compared to the reactor volume, so the concentration of dye can be assumed to be constant. The exit from the reactor leads to a flow cell where the UV–VIS transmission of the stream is measured. After the measurement, the stream is led to waste. The reactor gives the system a large dead volume that retains the injected pectin, which is only slowly washed away by the carrier stream, see Fig. 2.

The dye is known to bind to the pectin irreversibly (within the time frame of the experiment) with a high affinity, and changes conformation (i.e. absorbance) depending on whether it is bound or not bound to the pectin. It is observed from initial studies, that the dye, when bound to the pectin, also changes conformation depending on whether the neighbouring galacturonic acid on the pectin backbone contains an ester group or an acid group (with another dye molecule attached), i.e. the dye conformation is different if the molecule binds to a blocky or to a random pectin environment.

The concentration of pectin injected is initially high compared to the concentration of dye in the reactor. It is therefore assumed that the dye binds preferably to the easily accessible sites on the pectin. As the pectin (with dye attached) is washed away from the reactor by the continuously flowing carrier stream (still containing the dye in a fixed concentration), the concentration of pectin decreases exponentially. As the concentration of pectin drops, the dye will also bind to the less accessible sites on the pectin. At the end of the sample run, the pectin concentration will be close to zero and negligible. Hence, only the unbound dye will be spectroscopically active at that stage, see Fig. 3.

2.3. Data collection

UV–VIS spectra were recorded in transmission mode using a StellarNet EPP2000 instrument with the SpectraWiz Spectrophotometer software version 3.3, which reported the measurements in

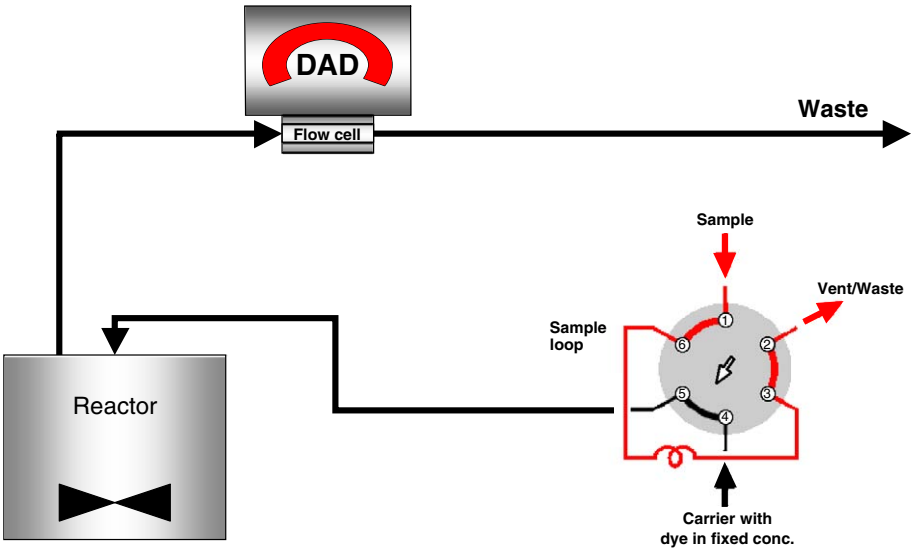


Fig. 2. Overview of the setup.

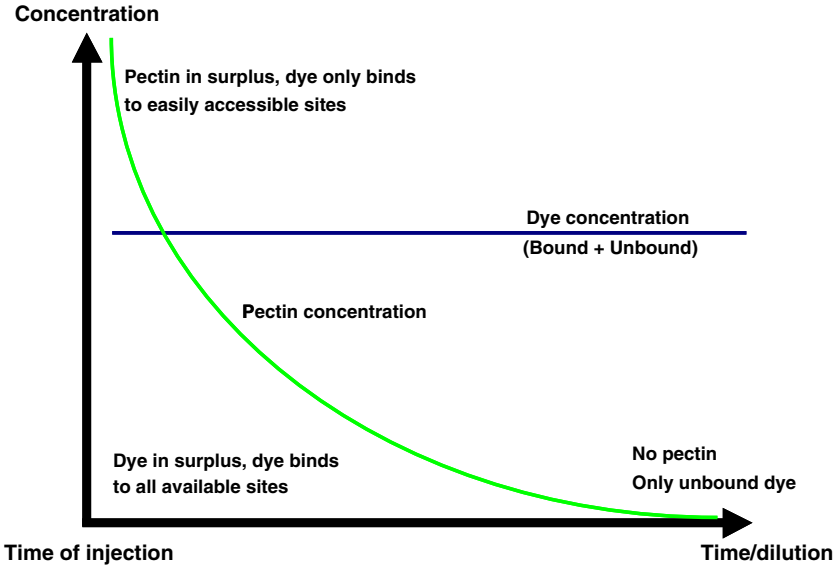


Fig. 3. Idealized concentration of pectin and dye in the reactor tank.

absorbance units. For every sample run, the raw data matrix collected consists of 333 time points and 2048 wavelengths. In order to facilitate the data analysis, the data was reduced by averaging wavelengths and time points and cropped by eliminating wavelengths with no relevant information. Time points at the start and the very end were also eliminated, as they contain artefacts of the injection and mixing of the sample. The reduced data set of all 33 samples (including the remethylated and the mother pectin) is sized 33 samples × 77 time points × 145 wavelengths.

The data was pre-processed using Matlab version 7.0.4.365 (R14 SP2) from The MathWorks and the PLS-Toolbox, version 3.5.2 from Eigenvector Research.

In order to compensate for small run-to-run variations (minor baseline drift and small offsets in dye concentration), the data was corrected by fitting a straight line through all times of the first and last wavelength (i.e. wavelength 1 and 145, as no absorbing components are seen here) of the individual sample landscape and subtracting the resulting square made by connecting these time points from the landscape. The corrected data matrix was then scaled to same intensity of wavelength number 55 (which is the peak of the largest dye peak) at the last time point where only unbound dye is observed. The correction did not result in large changes in the overall appearance of the landscapes, but did result in improvements in the computed lack of fit.

Models were calculated using the MCR-ALS with the graphical user interface, version 1.0.0 [15] and PARAFAC2 [16]. The MCR-ALS algorithm was modified, to accommodate for automated input of similar matrix size for all sample runs, which is not implemented in the published version of the MCR-ALS algorithm [17]. The PARAFAC2 algorithm was modified by the authors to disable the conventional automatic sorting of the components according to largest variance explained.

Reasons of confidentiality prevent the publication of pertinent details of the setup such as i.e. specific flow rates, absolute wavelengths and times, volumes and the dye involved.

2.4. Overview of selected samples

Fig. 4 shows 16 landscapes out of the 33 available samples in the data set. The landscapes are presented as 2-D contour plots with time on the *x*-axis and wavelength on the *y*-axis. Absorbance is represented by equal-intensity contours in the matrix. The landscapes are arranged so as to reflect the sample design.

Visual inspection and initial modeling reveal that three main phenomena are present. 1) A peak identified as a random fingerprint appears at wavelength range 70–90, time 1–40. 2) A peak identified as a blocky fingerprint appears at wavelength 28, time 1–40 and 3) A fingerprint identified as the unbound dye returning in the system, is identified as a double peak at wavelengths 55 and 75 appearing between times 1 and 60, and always continuing until the end of the sample run. The end spectrum in all samples match the spectrum of the carrier with dye with no pectin added, and the later the unbound dye appears in the spectra, the more deesterified the sample is. The "Remethylated" pectin has only the profile of the unbound dye, thereby indicating that even at high pectin concentration, the dye does not bind to anything but the galacturonic acid groups on the pectin.

The random profile does not shift in time. Intensity is dependent on the amount of randomness introduced (high amount= high intensity). Some randomness is also introduced by blocky deesterification, see e.g. sample "B36.8 R0.0". The random spectral profile appears not to be well defined. The spectrum is non-Gaussian as it is very wide, which could suggest several underlying components or overlapping peaks forming the peak.

The unbound dye is made up of two peaks (wavelengths 55 and 75). Sometimes, the wavelength 75 peak appears "before" the wavelength 55 peak. This could also be due to an additional or intermediate compound appearing as part of the random profile.

The blocky profile always appears after the random profile, and the more blockiness introduced, the later the blocky profile peaks. A high degree of "time-shift" in the blocky profile is

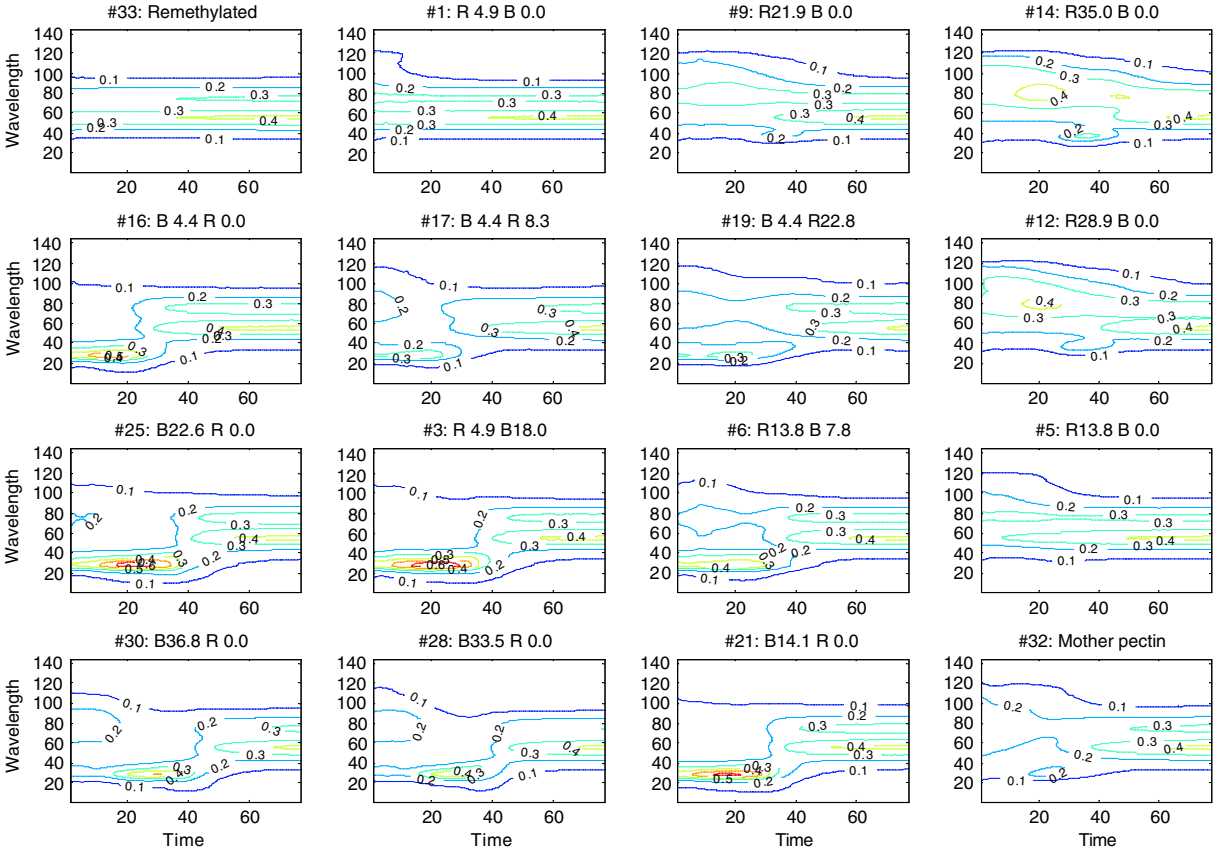


Fig. 4. Landscapes for 16 of the 33 samples. Absorbance is represented as contour lines. The landscapes are arranged with the remethylated sample in the top left corner and the most blockwise and random deesterified samples in the lower left corner and top right corner, respectively.

present in the system and the time-shift seems dependent on the amount of randomness present, as the random profile always appears before the blocky profile. The highest intensity (over all 33 samples) is observed in sample "B22.6 R0.0". Introduction of randomness weakens the blocky profile. Some blockiness is also introduced by random deesterification, as seen in sample "R35.0 B0.0".

The unbound dye returning also exhibits a strong time-shift, dependent on the degree of deesterification as mentioned. At the end of the blocky profile, a small spectral shift from wavelength 28 to 33 is observed (See, for instance, "R4.9 B18.0"). This is most likely due to an additional or intermediate compound, rather than a change in physical environment (pH, ionic strength, etc., that can affect the electronic states of the absorbing species, are controlled during the experiment).

The presence of time-shifts and other shape variations in the time profiles, but also the possible additional or intermediate compounds present, makes MCR-ALS and PARAFAC2 good choices for fitting the data. If the data set is well resolved, it should be possible to quantify the amount of blockiness and randomness from the identified concentration profiles, provided that some sort of scale or index is defined. The correlation will not necessarily be straightforward though, as deviations from linearity are observed e.g. between concentration/amount of the blocky profile and the blockiness. An article quantifying the amount of blockiness and randomness based on the results presented in this paper is submitted by the authors [14], and the results were presented at the 9th Scandinavian Symposium on Chemometrics, Reykjavik, Iceland, August 2005.

3. Methods

3.1. Multivariate curve resolution

Multivariate curve resolution (MCR) is a commonly used technique that can resolve multi-component mixtures into a simple model consisting of a composition-weighted sum of the signals of the pure compounds [9,12,18]. The multivariate curve resolution model can be written as:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

It is a bilinear method to resolve an experimental data matrix $\mathbf{D}$ ($I \times J$) into the product of a column matrix $\mathbf{C}$ ($I \times F$) usually associated with concentration profiles, and a matrix of row profiles $\mathbf{S}^{\mathrm{T}}$ ($F \times J$), usually associated with spectra. The matrix $\mathbf{E}$ ($I \times J$) is the residuals i.e. what is not explained by the model $\mathbf{C}\mathbf{S}^{\mathrm{T}}$. The entries in matrix $\mathbf{E}$ should be small and random compared to the numbers in $\mathbf{D}$ and the model $\mathbf{C}\mathbf{S}^{\mathrm{T}}$. The scalars I and J are the total number of time points in the data set studied and wavelengths of the data set, respectively, and F is number of components to be resolved.

This general description can fit a wide range of applications, processes, mixtures, elutions, images and environmental data monitored by multivariate instrumentation [13,18].

The model parameters are estimated using an Alternating Least Squares (ALS) algorithm that iteratively fits the $\mathbf{C}$ and $\mathbf{S}^T$ matrices to the experimental data $\mathbf{D}$. The model is fitted with a pre-defined number of components, F, using initial estimates of either the $\mathbf{C}$ or the $\mathbf{S}^T$ matrix. Exploratory data analysis using Singular Value Decomposition, Evolving Factor Analysis (EFA) [19–21] or SIMPLISMA [22] can provide knowledge used for the initial estimates.

During the MCR-ALS iterations, it is possible to impose constraints such as non-negativity, unimodality, closure, trilinearity and/or selectivity [9]. The use of constraints can facilitate the convergence of the algorithms, resolve rotational ambiguities and ensure that the computed solutions are chemically meaningful [23]. The number of components and use of constraints will be further elaborated upon later in this paper.

A powerful extension of MCR-ALS is multi-sample MCR-ALS, where it is possible to resolve several independent samples and/or several independent measurement techniques by augmenting the input matrix appropriately [9,12,13]. Extending the concentration matrix column-wise to estimate concentration profiles of several samples will facilitate the resolution of a global solution and can solve the resolution of components in situations of rank-deficiency [18].

In a column-wise augmented data set, the matrices in the Eq. (1) model must be interpreted as: $\mathbf{D}=[\mathbf{D}_1;\mathbf{D}_2;\ldots;\mathbf{D}_K]$ and $\mathbf{C}=[\mathbf{C}_1;\mathbf{C}_2;\ldots;\mathbf{C}_K]$ with K being the number of samples analysed simultaneously (In MATLAB notation, the semicolon stands for appending matrices one on top of each other). $\mathbf{S}$ is a single matrix with pure resolved spectra common to all samples and $\mathbf{D}_k$ and $\mathbf{C}_k$ are the raw data and the resolved concentration profiles for the k^{th} sample. Because of the column-wise data arrangement, the number of rows in $\mathbf{D}_k$ and the shape of the related concentration profiles in $\mathbf{C}_k$ can show any kind of variation from sample to sample.

In this study the concentration mode will reflect the dilution profiles of the sample matrix. This way the algorithm will resolve common pure spectra ($\mathbf{S}$) for all samples and will identify individual time/concentration profiles ($\mathbf{C}_k$) for each sample. For every time point in every individual sample, the "amount" of every pure spectrum (constituent) will be quantified. This will be a relative number, depending on the sample and dilution, but as the dilution is kept the same from sample to sample, they are comparable. This relative number is still referred to as concentration throughout the article, but should not be thought of as an absolute concentration.

3.2. PARAFAC2

PARAFAC2 is derived from the PARAFAC algorithm, an algorithm that models a three-dimensional array $\underline{\mathbf{X}}$ ($I\times J\times K$) as a summation over outer products of F triads of vectors [24]. The scalars I, J and F still hold the number of time points and wavelengths of the data, respectively, and the number of components to resolve. The scalar K holds the number of samples.

PARAFAC2 is an advanced variant of the PARAFAC algorithm [10] developed to handle the situation where the number of observations in one mode varies or where shifts or shape-changes of profiles along one mode are anticipated. This makes the algorithm appropriate to handle, for instance, retention time shifts [11]. It is investigated in this article whether the algorithm is able to handle sample differences in the resolved time/concentration profiles.

The matricised notation of the PARAFAC2 model can be written as:

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k(\mathbf{P}_k\mathbf{H})^T + \mathbf{E}_k, \quad k = 1,..,K \tag{2}$$

Where $\mathbf{X}_k$ represents the data related to one sample (one slab from the original data cube) of size ($J\times I$) in which I can vary with k. K is the number of samples. $\mathbf{A}$ ($J\times F$) holds the first-mode loadings, in this study the resolved spectra. $\mathbf{D}_k$ ($F\times F$) is a diagonal matrix that holds the k'th row of $\mathbf{C}_{P2}$ in its diagonal. The notation $\mathbf{D}_k$ is maintained for consistency with the literature, but note that $\mathbf{D}_k$ is not related to the $\mathbf{D}_k$ used in the MCR model. $\mathbf{C}_{P2}$ ($K\times F$) holds the third mode loadings, in this study the sample scores. $\mathbf{H}$ is an $F\times F$ scaling matrix, and $\mathbf{P}_k$ is an $I\times F$ orthonormal matrix (I can vary with k). The matrix $\mathbf{E}_k$ holds the residuals. $\mathbf{P}_k$ and $\mathbf{H}$ have no direct chemical or physical interpretation but their product will be an estimate of the time profiles.

The chosen notation is not completely transparent, but is motivated by an effort to match as closely as possible the choice of notation in the algorithms used. It should be especially noted that the $\mathbf{X}_k$ mentioned in PARAFAC2 reflects the transpose of the data matrix $\mathbf{D}_k$ in MCR-ALS and that the $\mathbf{C}_{P2}$ from PARAFAC2 is not equivalent to the $\mathbf{C}$ resolved from MCR-ALS.

In order to compare the profiles resolved by MCR-ALS and PARAFAC2, the resolved spectra $\mathbf{S}^T$ from MCR-ALS in Eq. (1) and $\mathbf{A}$ from PARAFAC2 are both scaled to unit length and then directly comparable. The concentration profiles from the individual samples can be computed from the PARAFAC2 model by:

$$\mathbf{C}_k = \mathbf{D}_k(\mathbf{P}_k\mathbf{H})^T, \quad k = 1,..,K \tag{3}$$

$\mathbf{C}_k$ can then be compared to the analogous submatrix of the augmented $\mathbf{C}$, as resolved by multi-sample MCR-ALS.

The PARAFAC2 model differs from a Principal Component Analysis (PCA) of the unfolded array $\underline{\mathbf{X}}$ by the requirement that the spectral loadings $\mathbf{A}\mathbf{D}_k$ must be proportional for every value of k (running from 1 to K). Furthermore, it is assumed that the cross-product $(\mathbf{P}_k\mathbf{H})^T(\mathbf{P}_k\mathbf{H})$ is constant for all values of k [10,25]. This condition is sufficient for the PARAFAC2 model to be unique if some other mild uniqueness conditions are met, as further discussed in depth in Kiers et. al. [10].

Even though the matricised notation of PARAFAC2 is presented, it should still be considered a multi-way method with an independent sample mode. In MCR-ALS, the sample mode is only implied through extension of the concentration matrix. This has implications on how constraints are imposed, as discussed in the next section.

In summary, the PARAFAC2 model is more flexible, compared to the PARAFAC model. It allows for deviations of the inherent trilinearity imposed by the PARAFAC model. The concentration profiles resolved on the time axis are allowed to vary as long as the before mentioned cross-product is constant over different samples. This does give the PARAFAC2 model properties comparable to the multi-sample MCR-ALS model, motivating the comparative study in this paper.

3.3. Use of constraints

As is revealed by data inspection and apparent from Fig. 4, the acquired absorbance spectra are positive in the whole time domain and it is reasonable to expect that the underlying spectra will be non-negative. Likewise, it is not chemically meaningful to estimate negative concentrations. Thus, imposing non-negativity constraints in the spectral and time mode can be helpful. In some systems, it is observed that the solution by itself converges to a non-negative solution without the constraints applied. But even then, the use of the constraint can affect the number of iterations needed for convergence significantly. Non-negativity can be applied in the algorithms in several ways. It is chosen to implement non-negativity in MCR-ALS by a fast non-negativity least squares algorithm [26], which is the same as the one used in the PARAFAC2 algorithm. Implementation of non-negativity in the spectral mode is straightforward and similar for both algorithms. There is however an important difference in how non-negativity is imposed in other modes. MCR-ALS resolves concentration profiles directly, and the non-negativity constraint can be imposed on them as such. PARAFAC2 does not resolve the concentration profiles, but rather sample mode loadings ($\mathbf{C}_{P2}$), a scaling matrix ($\mathbf{H}$) and an individual sample specific orthonormal basis ($\mathbf{P}_k$). Non-negativity cannot be imposed on the time profiles, $\mathbf{P}_k\mathbf{H}$, for algorithmic reasons. Non-negativity constraints can only be imposed on the sample mode loadings and this can still lead to negative concentration profiles. A direct comparison of non-negativity constrained MCR-ALS and PARAFAC2 models will therefore not be possible.

Unimodality is another constraint that is relevant to consider. Unimodality means that the profile (concentration or spectrum) has one maximum and is monotonically increasing before that maximum and monotonically decreasing after the maximum. The maximum does not have to be present — if the function is only monotonically increasing or decreasing in the observed range, the criterion of unimodality is still fulfilled. A unimodality constraint tolerance parameter in MCR-ALS allows local departure from the unimodality. A tolerance of e.g. 1.05 means that departure of 5% of the value of the main peak is allowed before the constraint is enforced [13] i.e. the constraint is made less strict. The constraint is applied in the "average" implementation in the MCR-ALS algorithm, which allows the profile to increase from a local minimum by a straight line to a local maximum provided the tolerance limit is not exceeded. The algorithm is based on a unimodal (and non-negativity) least squares algorithm [27]. This is similar to the implementation in the PARAFAC2 algorithm, where however no tolerance parameter is available.

As discussed in the instrumental setup, the concentration of pectin decreases exponentially, whereas the dye is kept at a fixed concentration. Therefore — if the system behaves in accordance with expectation — the observed concentrations should increase or decrease exponentially. But as can be seen from Fig. 4 this is not the case. Both the random and the blocky peaks are observed to be first increasing, then decreasing in intensity. It might thus be concluded that the system is not as simple as expected. This could be due to relatively slow reaction kinetics. As the concentration of pectin decreases, the dye will start to bind to the less accessible (i.e. blocky) sites on the pectin and the concentration of the blocky fingerprint will increase. This however still complies with the unimodality constraint.

The MCR-ALS algorithm offers the possibility to impose trilinearity for one or more components. This means that the resolved components in different samples have the same shape. If imposed for all components, the MCR-ALS model will approach the PARAFAC model [24]. Since the concentration profiles are thought to be sample dependent, no assumptions about trilinearity are made in this study.

In summary: in this paper results using non-negativity and unimodality constraints are presented. For visualization purposes, resolved spectra are normalized to unit length in both the PARAFAC2 and the MCR-ALS algorithms.

3.4. Number of components

As discussed previously, three overall phenomena or fingerprints can be visually identified within each sample: the randomly distributed acid groups, the blocky sequences of acid groups and the appearance of the unbound dye as the pectin is diluted (see Fig. 4). Therefore, at least three components are expected to be needed in the models. Since the spectral mode is common for all the samples, any spectral shift will require further independent components to be modeled.

Preliminary exploratory data analysis using Singular Value Decomposition (SVD) and Evolving Factor Analysis (EFA) [19–21] combined with data inspection revealed that appropriate models could be made using three to five components. Models with two and six components were computed, but more than five components led to clearly overfitted solutions. Therefore, results using three to five components will be investigated in this article.

3.5. Initialisation of algorithms

In order to ensure comparability of the computed solutions, both algorithms were initialised with a set of up to five spectra computed from preliminary MCR-ALS sample runs (see Fig. 5). The spectra were manually inspected and edited using experience gained during exploratory studies. Some peaks that appeared ambiguous or inconsistent were removed from two of the spectra (the Blocky2 and Blocky + Random component) used for initialisation.

The risk of using a fixed initialisation scheme is that one cannot be certain that the global solution with the best fit will be found; rather the algorithms could converge in a local minimum

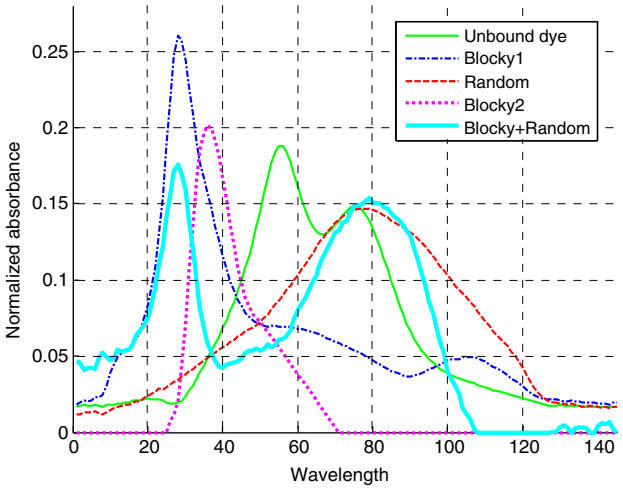


Fig. 5. Spectra in selected order (e.g. a three component model will be initialised with the first three spectra) used for initialisation. The spectra are edited from initial MCR-ALS runs. The spectral labels reflect the chemical/physical interpretation of the authors.

in terms of fit. This is checked by — in selected examples — repeating the algorithms many times, using random values for initialisation and checking that the loss function values are similar or higher. All solutions were inspected visually, and it was confirmed that all resolved spectral and concentration profiles were similar in a meaningful way, indicating that all computed models converged to a related global solution, which can be compared regardless of number of components identified.

In the MCR-ALS toolbox, convergence is achieved when the change of the standard deviation of the residuals between two consecutive iterations is less than 0.1%. In the PARAFAC2 algorithm, the convergence criterion is an absolute change lower than $1*10^{-6}$ in percent variance explained. In both cases the convergence criterion was sufficient to ensure no significant change in the solution.

4. Results and discussion

The computed solutions of the algorithms will be presented in terms of the lack-of-fit in percentage plus the resolved spectral and concentration profiles. The lack of fit relates to the difference between the squared sum of the input data and the squared sum of the modeled variation with the resolved PARAFAC2 or MCR-ALS profiles. For one sample it is computed according to the following expression:

$$\text{lack of fit}(\%) = 100\sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \quad (4)$$

where d_{ij} is an element in the raw data (offset corrected) and e_{ij} is the corresponding element in the residuals after the modeled variation is removed. A low lack of fit percentage indicates that a model fits the data well.

Table 2 lists the computed lack of fit summarized for all samples for models run under selected conditions. MCR-ALS models are run both unconstrained, and with unimodality imposed (in the concentration mode only) with a tolerance of 5% and 0%, respectively. Both MCR-ALS models with unimodality imposed also have non-negativity imposed (in both the spectral and concentration mode). The PARAFAC2 models are run either unconstrained or non-negativity constrained in both the spectral mode and the sample mode.

Overall, the lack of fit is low, so all computed models describe the data well. The approach of fitting the same spectral profiles and individual concentration profiles to the 33 samples seems to work well for both algorithms.

The MCR-ALS models with non-negativity constraints imposed do not lead to an increased lack of fit compared to the unconstrained MCR-ALS models. Thus, the non-negativity restriction is not actively constraining the solution, but merely restricting the solution-space. The PARAFAC2 models with non-negativity imposed on just the spectral mode converge to the same solutions as the unconstrained models, as the unconstrained solutions are non-negative by itself in the spectral mode. For this reason the results are not repeated in Table 2. The PARAFAC2 models with non-negativity imposed on both the spectral mode and sample mode do not lead to credible solutions as non-negativity imposed on the $\mathbf{C}_{P2}$ matrix in PARAFAC2 heavily distorts the **P** and **H** matrix. This worsens the fit and gives chemically unlikely concentration profiles for all number of components resolved.

As the lack of fit only differs by a few percent, no conclusions regarding the performance of MCR-ALS compared to PARAFAC2 can be drawn from Table 2 alone, but the higher lack of fit of PARAFAC2 models could be related to the allowed changes in the time profile shapes as imposed by the constant cross-product constraint in the PARAFAC2 model, as opposed to the complete freedom allowed in the modeling of the time profile shapes by MCR-ALS. This can be revealed by close inspection of the computed models.

Resolving six components seems to be clearly overfitted judged from the appearance of the resolved spectra and concentration profiles. Chemically improbable concentration profiles and spectra, which are almost identical to spectra of other components, are resolved.

The unimodality constraint imposed on the MCR-ALS concentration profiles does not lead to significant changes in the resolved spectra. Rather, the unimodality constraint is seen to cause artifacts on the resolved concentration profiles. By inspection, the most promising solutions seem to be the three

Table 2
Overview of computed lack of fit (%) of selected models computed

Lack of fit (%)	3 comp.	4 comp.	5 comp.	6 comp.
MCR-ALS, unconstrained	5.17	3.76	2.43	2.03
MCR-ALS, non-negativity	5.17	3.76	2.43	2.03
MCR-ALS, unimodality 1.05	5.29	4.76	3.76	4.29
MCR-ALS, unimodality 1.00	5.41	4.92	3.76	4.04
PARAFAC2, unconstrained	8.30	5.01	4.13	3.66
PARAFAC2, non-negativity	10.50	5.44	5.06	3.91

The MCR-ALS unimodality constrained (in concentration mode only) models are non-negativity constrained in both the concentration and spectral mode.

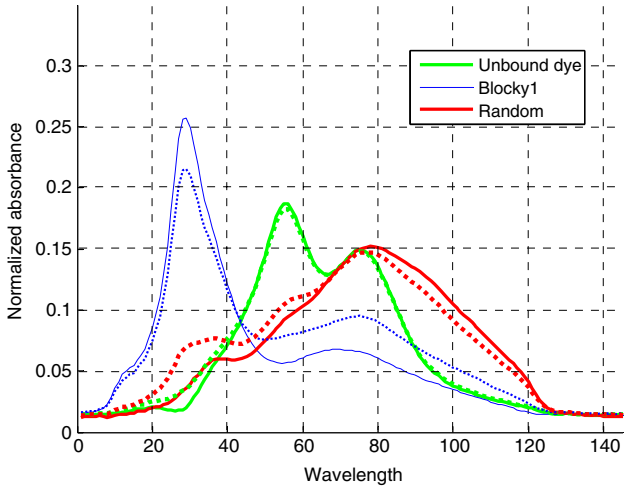


Fig. 6. Resolved spectra for unconstrained three component models. MCR (solid lines) and PARAFAC2 (dotted lines).

and five component models. The four component models seem to have a fourth component with features representing a mixture of two components. It appears that the variation of the fourth component is better described by two components with independent concentration profiles. As the unimodality constrained models with a tolerance limit of 1.00 is stricter than 1.05, an increase in the lack of fit is seen for the three and four component models. The lack of fit is increasing for the unimodality constrained six component models compared to the five component models. This is an indicator of a local minimum in the solution and, indirectly, overfitted models.

Results from unconstrained MCR-ALS and PARAFAC2 models with three and five components will be presented in greater detail. Even though non-negativity constrained (without unimodality) MCR-ALS solutions fit equally well and make more chemical sense, the results from the unconstrained solution are presented in this article in order to compare with equivalent PARAFAC2 solutions. Results using the MCR-ALS solutions with non-negativity constraints imposed, are used for quantifying the amount of blockiness and randomness induced in the pectin, and is presented in Zachariassen et. al. [14].

The three component models provide information about the unbound dye returning to the system, a component attributed to the presence of blocks (the Blocky1 profile) and a component attributed to the presence of random deesterification (the Random profile). This is consistent with the results of visual inspection presented in the overview section.

An inspection of the square root of the average squared sample residual landscapes (not shown) reveals that the three component MCR-ALS, and PARAFAC2 model in particular, leaves variation of the random and blocky fingerprint unexplained, whereas the 5 component models, the MCR-ALS models in particular, account for most systematic variation leaving average sample residuals less than 0.01 absorbance units. All models are capable of fitting the unbound dye returning in the system. The unexplained variation of the blocky fingerprint of the MCR-ALS three component model can be attributed to the observed wavelength shift,

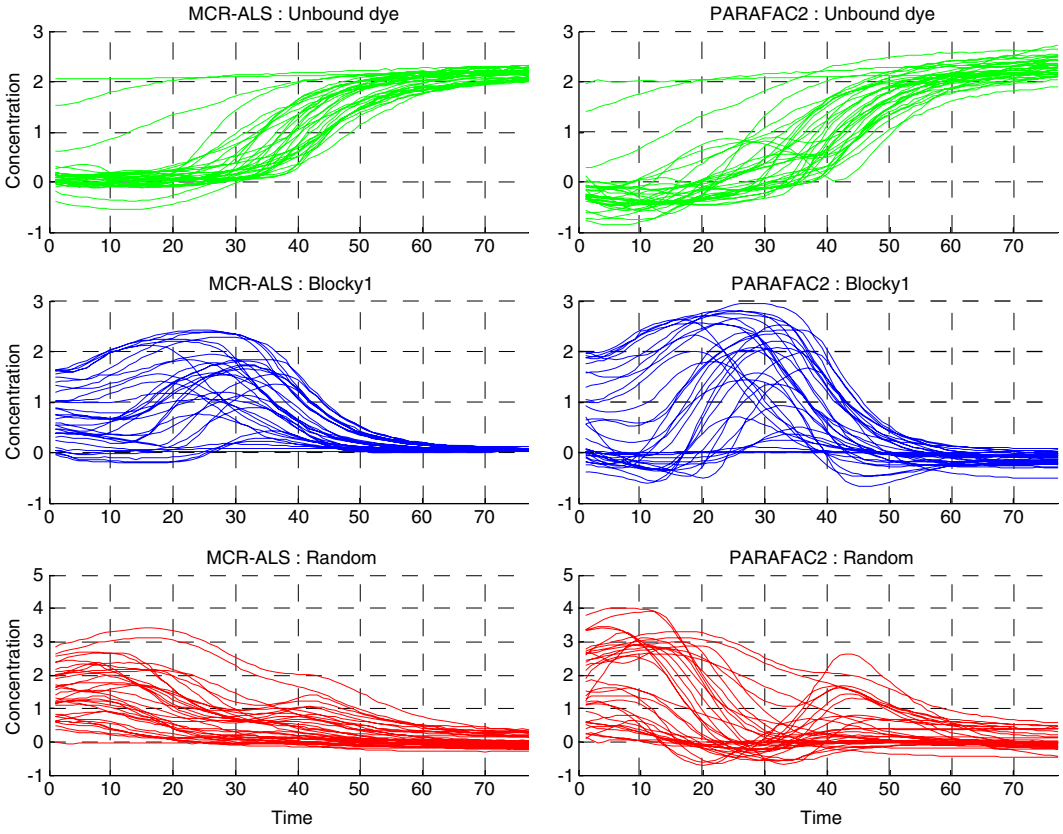


Fig. 7. Resolved concentration profiles for unconstrained three component models. MCR-ALS profiles are to the left and PARAFAC2 profiles are to the right.

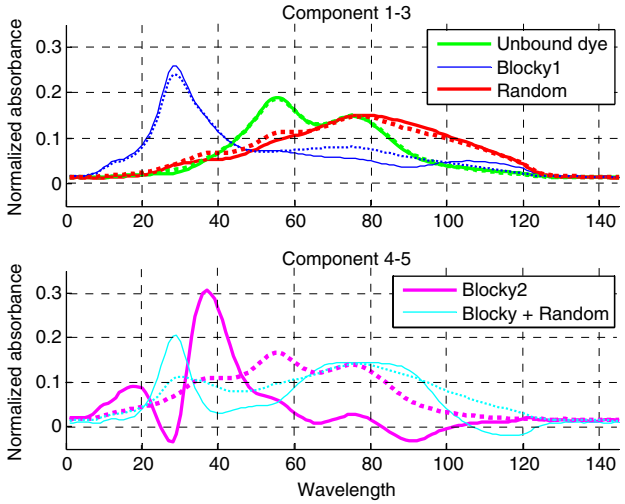


Fig. 8. Resolved spectra for 5 component models, unconstrained. MCR (solid lines) and PARAFAC2 (dotted lines). The Blocky2 PARAFAC2 component is very similar to the Unbound dye, while the Blocky2 MCR-ALS component models a wavelength shift of the Blocky peak.

whereas the PARAFAC2 three component model miss more general aspects of the fingerprint as well.

In Fig. 6 the resolved spectral profiles for the unconstrained three component PARAFAC2 and MCR-ALS models are given. The profile for the unbound dye is quite similar, while the Blocky1 and Random profiles have slightly different appearance.

In Fig. 7, the resolved concentration profiles for the unconstrained three component models can be seen. The unbound dye has a sigmoidal profile. It is close to zero at the beginning of the dilution where all the dye is supposed to be bound to the pectin. Later, when the concentration of pectin is decreasing, the resolved concentration profile of the unbound dye increases to about two units. The time where the concentration profile of the unbound dye starts to increase is related to the total deesterification of the pectin. The more deesterified the pectin, the later the profile of the unbound dye starts to increase. Three samples have a high resolved concentration profile already from the start of dilution. They are the profiles of samples “R13.8 B0.0”, “R4.9 B0.0” and the remethylated pectin, which are all the samples with a very low degree of deesterification.

The Blocky1 profile initially has a low or intermediate resolved concentration, after which it increases and then drops towards zero. Both the time this phenomenon starts to increase, and the level it increases to, are dependent both on the amount of blockiness and randomness induced by deesterification. The overall relationship is not straightforward, but the more random the sample is, the later the Blocky1 profile appears and the lower relative intensity it will have.

The random profile initially has a rather high resolved concentration. For some samples there is a tendency for the concentration to modestly increase again around time point 40. The bimodality is more pronounced in the concentration profiles resolved by PARAFAC2. This is unexpected as the concentration of pectin is decreasing exponentially.

Overall, the concentration profiles resolved by unconstrained MCR-ALS tend to be close to non-negative with most concentration profiles positive or close to zero. The concentration profiles resolved by PARAFAC2 have higher negative values for longer time stretches. This is balanced out with higher resolved concentration values of the Blocky1 and Random component compared to the profiles resolved by MCR-ALS. This way negative concentrations of for instance the Blocky1 profile can be balanced by high positive concentrations of the Random profile thus keeping the overall fit to the recorded landscape good.

The addition of two components, for a total of five components — see Fig. 8, does not change the resolved spectra significantly for the first three components identified in the three component solutions (see Fig. 6). The two remaining minor components account for unexplained features of both blocky and random pectin deesterification patterns, as will be described later. It is reasonable that minor contributions can be of interest in the description of this system. Pectins are macromolecules and, as such, may have parts that do not fit exactly a completely random or completely blocky pattern, as described in the introduction. These small blocks or not fully random structures may be the source of the unexplained spectral features that are described with models with a higher number of components. Thus, the fourth component resolved by MCR-ALS has a peak at wavelength 37 and is named Blocky2. It models a wavelength shift in the Blocky profile from wavelength 28 to 37.

The resolved concentration profiles can be seen in Fig. 9. The MCR-ALS resolved concentration profiles for the first three of the five components are very similar to the profiles resolved in the three component solution seen in Fig. 7. This is not to the same degree true for the concentration profiles resolved by PARAFAC2. The fourth component resolved by MCR-ALS, the Blocky2 profile have overall low values, but some samples have an increase from time 30–40. As indicated by the resolved spectra, this reflects a slight wavelength shift on the blocky fingerprint. The increase is always seen in the samples immediately after the maximum of the Blocky1 profile. The fifth component resolved by MCR-ALS is tentatively named “Blocky+Random” and has a peak at the same wavelength (28) as the Blocky1 component, as well as a wide peak at wavelength range 70–85 which is similar to the random component. The resolved concentration profiles have low-intermediate values both positive and negative from times 1 to about 50 where after it decreases to zero. The spectrum of the resolved fourth component from PARAFAC2 looks like the unbound dye. It is a sign of overfit when PARAFAC2 resolves almost identical components. The fifth component resolved by PARAFAC2 resembles the fifth component resolved by MCR-ALS. The concentration profiles for the five PARAFAC2 components appear more distorted than the profiles resolved by MCR-ALS. All components except Blocky1 have bimodal concentration profiles with negative and large positive values, and small erratic fluctuations are observed.

In order to evaluate the similarity of the resolved individual sample-to-sample concentration profiles, the average correlation between the component-wise individual 33 sample concentration profiles resolved by MCR-ALS and PARAFAC2 are presented in Table 3 together with the standard deviation. A

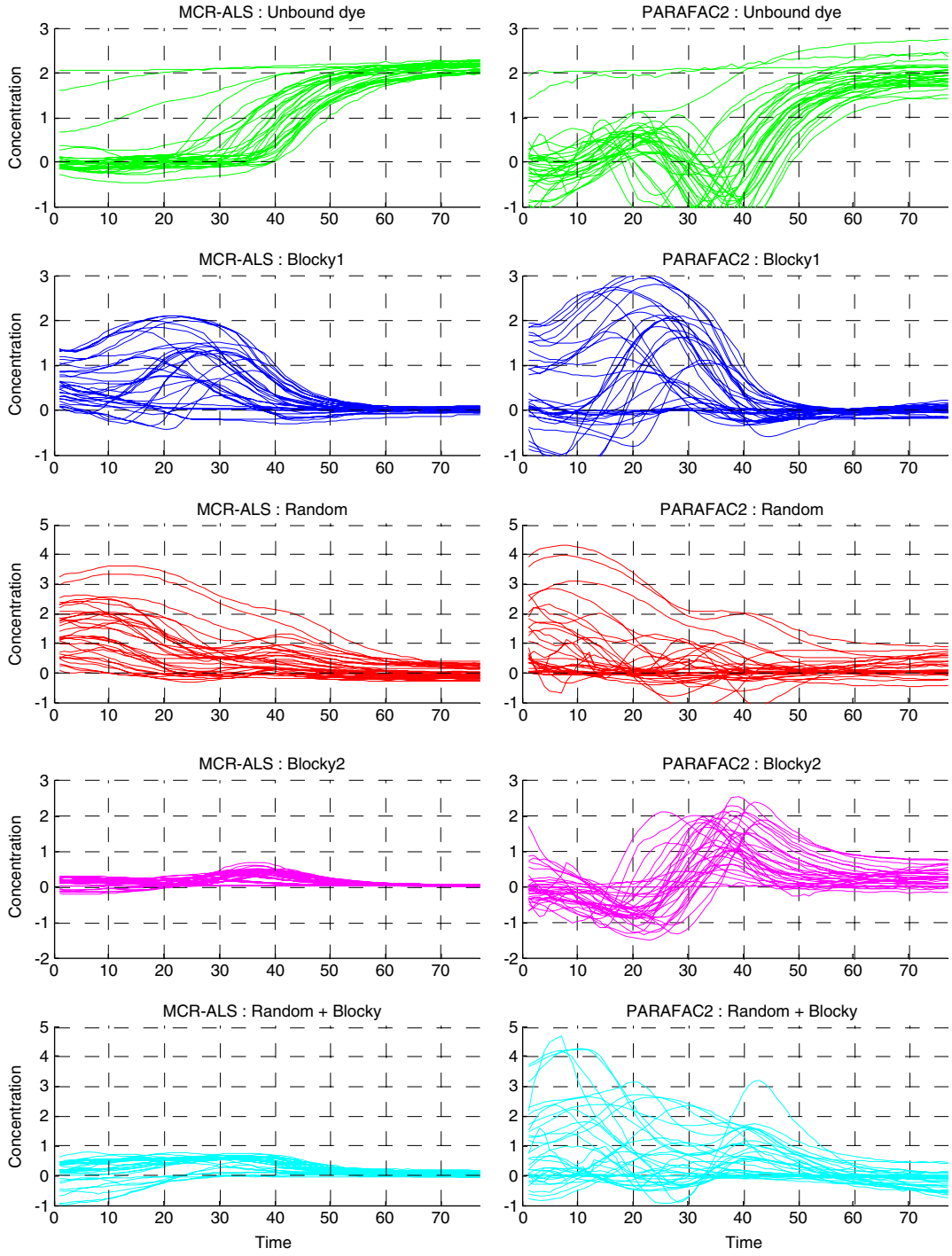


Fig. 9. Resolved concentration profiles for unconstrained five component models. MCR-ALS profiles are to the left and PARAFAC2 profiles are to the right.

high average correlation and a low standard deviation indicate that the shape of the resolved concentration profiles for that specific component, are very similar.

It is seen that for the three-component models, the concentration profiles are very similar, as high as 0.988 for the unbound dye. The component most differently resolved is the

Table 3
The average and standard deviation of the component-wise correlation between the concentration profiles of the 33 samples resolved by unconstrained MCR-ALS and PARAFAC2 models

Correlation	Unbound dye	Blocky1	Random	Blocky2	Random + Blocky
3 Comp Models	0.988 (0.021)	0.932 (0.115)	0.833 (0.175)		
5 Comp Models	0.856 (0.083)	0.753 (0.439)	0.490 (0.431)	0.615 (0.186)	0.320 (0.412)

Random with an average correlation of 0.833. Within the overall agreement of both methods, it is interesting to note the relationship between the sorting of the correlation coefficients and the shape of the related components. In a simplified manner, it can be said that the unbound dye has a simple sigmoidal shape, mainly affected by sample-to-sample shifts; the blocky profile has a unimodal shape that can shift and change width among samples and the random profile has a more irregular shape, sometimes showing more or less evident signs of bimodality in some of the samples. This matches with the fact that the dye and the blocky profiles look most similar among both resolution methods, whereas the random profile show a lower correlation coefficient and also a wider spread in the values from sample to sample. This could be explained by the higher ease of PARAFAC2 to recover profiles with more regular shifts or shape-changes, as those of the dye and the blocky profile.

For the five component models, the results are more dissimilar. Correlations are between 0.320 for the Random +Blocky profile and 0.856 for the unbound dye. Even though the resolved spectra visually judged are quite similar, the concentration profiles are different. This is due to the prevalent bimodality of many concentration profiles resolved by PARAFAC2 and the seemingly unstable high positive values needed to balance out negatively resolved concentration profiles. The minor contribution of the two added components and the less regular shape of their profiles (particularly for the fifth component) could explain the distorted PARAFAC2 results.

5. Conclusion

MCR-ALS and PARAFAC2 algorithms have been applied successfully to model the dilution profiles of 33 pectin samples in total. In order to compare the PARAFAC2 model with the MCR-ALS model, originally developed to model individual samples, the concentration matrices for MCR-ALS are extended or unfolded column-wise to estimate concentration profiles of several samples with a set of fixed spectral profiles. This makes the models directly comparable.

The computed lack of fit of the models varies between 2.0% and 8.3% using three to six components, when appropriate constraints are imposed. The lack of fit combined with inspection of the resolved profiles indicate that models with three or five components seem most appropriate.

Both the PARAFAC2 and the MCR-ALS algorithms successfully resolve three components, which can tentatively be attributed to three phenomena, viz. fingerprints of blocky acid groups, random acid groups and the unbound dye, the latter is seen when the concentration of pectin is diluted to a sufficiently low concentration. The three component PARAFAC2 and MCR-ALS models are very similar both in terms of the common spectral profiles resolved and the individual sample concentration profiles.

An unconstrained MCR-ALS gives the lowest lack of fit since it allows for the resolution of the concentration profiles in a more flexible way. This allows for two minor components further to be resolved describing minor spectral wavelength shifts and/or artefacts observed at very low pectin concentrations when the unbound dye starts to increase in concentration. The additional components cannot be resolved fully by PARAFAC2 where signs of overfit are observed.

It is difficult to directly compare other models than the completely unconstrained models, because non-negativity constraints are implemented differently in the algorithms.

The unconstrained three component models lead to “near” non-negative solutions for all components, they are simple and parsimonious in regard to the investigated problem, and the fact that PARAFAC2 and multi-sample MCR-ALS give an almost similar result add to the credibility of the models. The method successfully resolves different components depending on the deesterification pattern induced in the pectins. Information with regard to amount and dilution time resolved in the concentration profiles can be related to the deesterification pattern of the pectins. The resolved models therefore provide a good starting point for further interpretation and quantification of the amount of blockiness and randomness introduced by the enzymatic deesterification of the pectin [14].

Acknowledgements

The authors would like to gratefully acknowledge the grant made available by the Ministry of Science, Technology and Innovation to Christian B. Zachariassen, to partly cover the expenses in connection with participation in the Industrial Ph.D. project, which is a joint effort between CP Kelco and the Quality and Technology Group, The Royal Veterinary and Agricultural University of Denmark.

At CP Kelco, many colleagues have contributed in taking this application from idea to reality. Special credit is given to the laboratory personnel, viz. Anette Møller, Britta Pedersen and Hanne Thulstrup Jensen.

References

[1] C. Rolin, in: G.B. Saymour, J.P. Knox (Eds.), Pectins and their Manipulation, Blackwell Publishing, Oxford, Great Britain, 2002, pp. 222–241.
[2] Compendium of Food Additives Specifications/add. 9, FAO Food and Nutrition Paper no. 52, Rome, Italy, 2001.
[3] C. Rolin, J. De Vries, in: P. Harris (Ed.), Food Gels, Elsevier, Amsterdam, The Netherlands, 1990, pp. 401–434.
[4] P.J.H. Daas, K.M. Hansen, H.A. Schols, G.A. de Ruiter, A.G.J. Voragen, Investigation of the non-esterified galacturonic acid distribution in pectin with endopolygalacturonase, Carbohydrate Research 318 (1999) 135–145.
[5] P.J.H. Daas, A.G.J. Voragen, H.A. Schols, Investigation of the galacturonic acid distribution of pectin with enzymes part 2 — characterization of non-esterified galacturonic acid sequences in pectin with endopolygalacturonase, Carbohydrate Research 326 (2000) 120–129.
[6] G. Limberg, R. Korner, H.C. Buchholt, T.M.I.E. Christensen, P. Roepstorff, J.D. Mikkelsen, Analysis of pectin structure part 3 — quantification of the amount of galacturonic acid residues in blocksequences in pectin homogalacturonan by enzymatic fingerprinting with exo- and endo-polygalacturonase II from Aspergillus niger, Carbohydrate Research 327 (2000) 321–332.
[7] H. Winning, N. Viereck, L. Nørgaard, J. Larsen, S.B. Engelsen, Quantification of the degree of blockiness in pectins using ^{1}H NMR spectroscopy and chemometrics, Food Hydrocolloids, submitted for publication.
[8] K.S. Booksh, B.R. Kowalski, Theory of analytical chemistry, Analytical Chemistry 66 (1994) 782A–791A.

[9] R. Tauler, A.K. Smilde, B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, Journal of Chemometrics 9 (1995) 31–58.
[10] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2 — Part I. A direct fitting algorithm for the PARAFAC2 model, Journal of Chemometrics 13 (1999) 275–294.
[11] R. Bro, C. Andersson, H.A.L. Kiers, PARAFAC2 — Part II. Modeling chromatographic data with retention time shifts, Journal of Chemometrics 13 (1999) 295–309.
[12] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems 30 (1995) 133–146.
[13] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemometrics and Intelligent Laboratory Systems 76 (2005) 101–110.
[14] C.B. Zachariassen, J. Larsen, F. van den Berg, R. Bro, A. de Juan, R. Tauler, Multi-way analysis for investigation of industrial pectin using an analytical liquid dilution system, Chemometrics and Intelligent Laboratory Systems, submitted for publication.
[15] http://www.ub.es/gesq/mcr/mcr.htm as of June 2005.
[16] http://www.models.kvl.dk/source/ as of June 2005.
[17] J. Jaumot, Personal communication, 2005.
[18] A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution, Analytica Chimica Acta 500 (2003) 195–210.
[19] M. Maeder, A.D. Zuberbühler, The resolution of overlapping chromatographic peaks by evolving factor analysis, Analytica Chimica Acta 181 (1986) 287–291.
[20] M. Maeder, Evolving factor analysis for the resolution of overlapping chromatographic peaks, Analytical Chemistry 59 (1987) 527–530.
[21] H.R. Keller, D.L. Massart, Evolving factor analysis, Chemometrics and Intelligent Laboratory Systems 12 (1992) 209–224.
[22] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Analytical Chemistry 59 (1991) 1425–1432.
[23] A.K. Smilde, R. Bro, P. Geladi, Multi-way Analysis with Applications in the Chemical Sciences, John Wiley & Sons, Chichester, Great Britain, 2004.
[24] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (1997) 149–171.
[25] B.M. Wise, N.B. Gallagher, E.B. Martin, Application of PARAFAC2 to fault detection and diagnosis in semiconductor etch, Journal of Chemometrics 15 (2001) 285–298.
[26] R. Bro, S. de Jong, A fast non-negativity-constrained least squares algorithm, Journal of Chemometrics 11 (1997) 393–401.
[27] R. Bro, N.D. Sidiropoulos, Least squares algorithms under unimodality and non-negativity constraints, Journal of Chemometrics 12 (1998) 223–247.

Paper III

Multi-way analysis for investigation of industrial pectin using an analytical liquid dilution system

Christian B. Zachariassen, Jan Larsen, Frans van den Berg, Rasmus Bro, Anna de Juan, Romà Tauler

Chemometrics and Intelligent Laboratory Systems, **84** (2006), 9–20

Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 84 (2006) 9–20
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Multi-way analysis for investigation of industrial pectin using an analytical liquid dilution system

Christian B. Zachariassen [a,b,*], Jan Larsen [a], Frans van den Berg [b], Rasmus Bro [b], Anna de Juan [c], Romà Tauler [d]

[a] Quality Control Development, CP Kelco ApS, Ved Banen 16, DK-4623 Lille Skensved, Denmark
[b] Department of Food Science, Quality and Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[c] Analytical Chemistry Department, Universitat de Barcelona, Martí i Franqués, 1-11, Barcelona, E-08028, Spain
[d] Department of Environmental Chemistry, IIQAB-CSIC, Jordi Girona 18, Barcelona, E-08034, Spain

Received 3 October 2005; received in revised form 2 March 2006; accepted 16 March 2006
Available online 3 July 2006

Abstract

Measurements from an analytical liquid dilution system are used to quantify the intra-molecular distribution of ester groups on the pectin carbohydrate backbone. Thirty-one pectins have been produced from remethylated pectin by enzymatically deesterifying them in steps to give different known ester distributions. The amount of deesterification is measured in each step by titration to provide the reference values. The system works by injecting a solution of pectin into a carrier stream containing a fixed concentration of dye. The dye binds site-specifically to the poly-α-(1→4)-D-galacturonic acids constituting the non-esterified parts of the pectin carbohydrate backbone. The carrier stream is led to a Continuously Stirred Tank Reactor (CSTR). The pectin is slowly diluted while UV–VIS spectra are recorded at the reactor outlet providing a landscape of wavelength-by-time for every sample. All pectins are measured this way in triplicate runs.

In an article preceding this, the acquired landscapes have been analysed qualitatively using Multivariate Curve Resolution (MCR) and PARAFAC2, both Alternating Least Squares (ALS) regression algorithms. It is concluded that the landscapes can be described by common spectral profiles for all pectins and individual concentration/time profiles for each sample run.

In this article, calibration to the reference values are done by multi-way Partial Least Squares (PLS) regression to correlate the acquired landscapes as independent variables directly to the reference values as dependent variables. Also, calibration is done by unfolding the landscapes for each sample run to a vector and use conventional PLS. The concentration/time profiles previously identified by MCR-ALS or PARAFAC2 are unfolded and used as independent variables in PLS rather than the whole landscape. Finally, the spectral information can be reduced even further by summing up or integrating the mentioned individual concentration profiles to just one number per profile identified by MCR-ALS or PARAFAC2 in the sample run, or using the identified score values for each profile from the PARAFAC2 model as new independent variables.

The most successful calibration models based on a calibration set built from one of the triplicates can predict the induced degree of deesterification to an error level of less than 3% in absolute values–corresponding to a 6% relative error to the calibration full range–when tested onto a set consisting of the two remaining sample runs from the full set. The best calibration models are based either on unfolded landscapes or unfolded concentration profiles resolved by MCR-ALS.
© 2006 Elsevier B.V. All rights reserved.

Keywords: PARAFAC2; Multivariate Curve Resolution; Pectin; Blockiness; Intramolecular distribution of esterification; Analytical liquid dilution system

1. Introduction

Pectin is a polysaccharide found in the primary cell walls and intracellular regions of higher order plants. It is a structural element giving both strength and flexibility to plant material. Pectin has excellent gelling, thickening and stabilizing properties and is widely used for commercial applications,

* Corresponding author. Quality Control Development, CP Kelco ApS, Ved Banen 16, DK-4623 Lille Skensved, Denmark. Tel.: +45 5616 5616; fax: +45 5616 9446.
E-mail address: christian.zachariassen@cpkelco.com (C.B. Zachariassen).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2006.03.010

primarily in the food industry [1–3]. Pectin mainly consists of a linear chain of poly-α-(1→4)-D-galacturonic acid with varying degrees of esterification (%DE). The distribution of the methyl ester groups on the pectin carbohydrate backbone is an important structural parameter that affects the pectin functionality in consumer applications.

The exact distribution of ester groups on the pectin backbone is a property that can only be defined for an individual pectin polymer, and this information is as such of no commercial interest as it is not possible to extract or produce identical pectin polymer molecules from natural sources. On the other hand, a method for measuring average intramolecular distribution of ester groups in bulk pectin is highly desired.

Different pectins extracted from natural lemon peel are investigated. In order to span the sample set beyond natural variation, 31 samples are prepared from a remethylated (i.e. esterified) pectin which is subsequently enzymatically deesterified by two different pectin esterases, known to deesterify pectin in either a predominantly "Random" or "Blocky" fashion. The terms Random and Blocky are loosely defined terms, but it has been suggested to define a blocky group as a sequence of four or more deesterified galacturonic acids next to each other in the pectin carbohydrate backbone [4,5], while Random groups in this context may cover truly random or more systematic (sequential) deesterification patterns, see Fig. 1.

Pectin esterase (sometimes referred to as pectin pectylhydrolase EC 3.1.1.11, pectin methylesterase, pectin demethoxylase, pectin methoxylase or pectase), de-esterifies the methyl esters of the carboxyl groups, producing methanol and low methylated pectin. The enzyme is produced by higher plants and microorganisms. Pectin esterases derived from plants [7,8] are thought to attack either at the non-reducing end or next to a free carboxyl group and to proceed along the molecule by a single chain mechanism, creating blocks of unesterified galacturonic acids that are very calcium sensitive [9]. Irregularities in the pectin galacturonan chain inhibit the activity of the pectin esterase. Irregularities could be acetylated monomers, the occurrence of carbohydrate side-chains (so-called hairy regions) or ester groups that are transformed into amides [7]. Pectin esterase is highly specific for the methylester of polygalacturonic acid. The rate of pectin deesterification depends on chain length; trimethyl trigalacturonate is not attacked at all [10]. Fungal pectin esterases differ from plant pectin esterases by obeying a multichain mechanism, removing methyl groups at random [11]. The enzymes used in this work for creating a blockwise pattern of non-methylated units is a pectin esterase found in a commercial enzyme preparation from papaya fruits. Properties of pectin esterase from papaya resembled those of pectin esterase retrieved from other plant sources [12]. The random pectin esterase is a commercial product called Rapidase FP Super® from DSM, The Netherlands.

How Blocky or Random pectin is, cannot be easily expressed by a single number. A method for quantifying these deesterification patterns in pectin will have to take into account that the average total degree of esterification (%DE) and the distribution of the ester groups will vary from sample to sample over a wide range, but also that the actual degree of esterification and the distribution of the ester groups will vary from molecule to molecule within a pectin sample. For this reason, it is proposed to determine the degree of esterification and the distribution of the ester groups by measuring Ultraviolet–Visible (UV–VIS) spectra of a solution of pectin at different time points in a system dynamically diluted with a solution of a fixed concentration of a dye, known to bind site-specifically only to the free acid groups of the pectin. This will provide averaged information from all pectin molecules in the solution. A Diode Array Detector (DAD) is used to acquire an entire spectrum for each time point. The spectra recorded will reflect the concentration gradient of the pectin as it is diluted, and every sample will provide a matrix with all the spectra of the pectin sampled. The matrix will have the format wavelength × time and it will contain the absorbance recorded. Such a system has the properties of a Continuously Stirred Tank Reactor (CSTR) and, analytically, will have features known from Flow Injection Analysis (FIA) and conventional High Performance Liquid Chromatography (HPLC), in the sense that the analyte is retained while subjected to chemical reactions. The data can be analysed taking advantage of the features linked to the analysis of multi-way data sets [13,14], which offer the potential to resolve the sought profiles uniquely and to provide quantitative information. In this sense, since the species associated with blockwise deesterification can be identified independently from random or sequential deesterification, the quantification can also separate both kinds of contributions, which will be very useful for total pectin characterization.

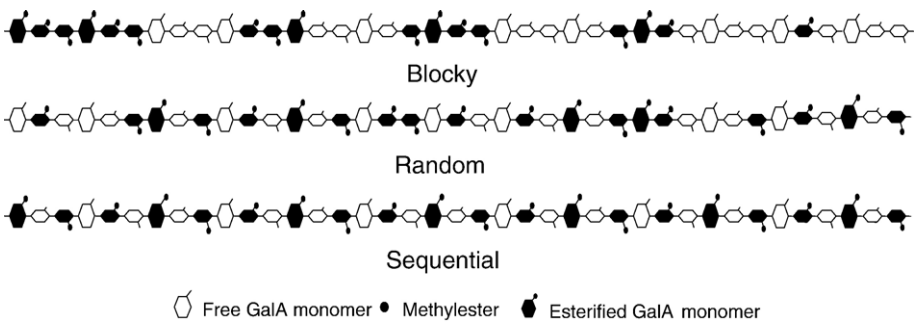


Fig. 1. Idealized diagram of three sections of pectin backbone (GalA = galacturonic acid). The pectins are deesterified 50% but the distribution of estergroups is different (modified from [6]).

The acquired wavelength × time landscapes can be calibrated to the reference values directly by N-way PLS or the landscapes can be unfolded by appending the acquired spectra to one vector per sample and apply conventional PLS. The landscapes can however also be modelled with common spectral profiles (for all samples) and individual concentration/time profiles using MCR-ALS or the PARAFAC2 model, which is a variant of the multi-way PARAFAC model, specially designed to handle shifts and/or individual sample or batch lengths in one of the modes, here the time which is synonymous with the concentration gradient mode. The resolved concentration profiles can then be unfolded and used for calibration by conventional PLS.

In an article preceding this [15] detailed resolution results, expressed as the pure spectra and time/dilution profiles of all dye-related species, i.e. free dye, and dye bound to blocky and random pectin were obtained, on one set of triplicate runs presented in this article, using and comparing the algorithms PARAFAC2 [16,17] and Multivariate Curve Resolution using Alternating Least Squares (MCR-ALS) [14,18–20]. The models were evaluated qualitatively with respect to fit and chemical interpretability and the applicability of the algorithms to this type of data were discussed in detail. The preceding article also discusses in greater detail the instrumental setup, data collection, the algorithms and constraints used, the appropriate number of profiles to extract and the qualitative interpretation of the resolved profiles. The reader is encouraged to read the article if a thorough understanding of the overall approach is required.

In this article, the acquired landscapes directly as well as the best models identified in the mentioned paper, is applied to quantify the degree and distribution of ester groups. Different approaches to the quantification will be demonstrated including use of the unfolded concentration profiles resolved by PARAFAC2 and MCR-ALS.

The same sample set has also been used in a quantitative study of the degree of blockiness using ^{1}H NMR spectroscopy [6], further validating the work presented in this paper.

2. Experimental

2.1. Sample preparation

The samples investigated are laboratory prepared materials derived from a natural lemon pectin–named the "Mother pectin"–produced at CP Kelco with a %DE = 72.3 and an average molecular weight of 180 kDa, determined by size exclusion chromatography (SEC). The pectin "Remethylated" was remethylated to %DE = 93.8 by taking 1027 g mother pectin and suspending it in 2000 mL methanol and 150 mL thionyl chloride at 0–5 °C for 5 days. The %DE was verified by titration using an internal CP Kelco control method. The standard deviation on the %DE titration method is 0.7% in absolute values. The average molecular weight was determined to be 13 kDa using SEC. This corresponds to a degree of polymerization of approximately 50 galacturonic acid units. The remethylated pectin is subsequently deesterified once or twice to give samples numbers 1–31 according to Table 1.

Table 1
Overview of the sample-set used in the present study

33: Remethylated pectin	*32: Mother pectin*			
1: R4.9 B0.0	*5: R13.8 B0.0*	*9: R21.9 B0.0*	*12: R28.9 B0.0*	*14: R35.0 B0.0*
2: R4.9 B10.6	6: R13.8 B7.8	10: R21.9 B9.7	13: R28.9 B13.2	15: R35.0 B6.8
3: R4.9 B18.0	7: R13.8 B13.8	11: R21.9 B21.6		
4: R4.9 B24.5	8: R13.8 B23.8			
16: B4.4 R0.0	*21: B14.1 R0.0*	*25: B22.6 R0.0*	*28: B33.5 R0.0*	*30: B36.8 R0.0*
17: B4.4 R8.3	22: B14.1 R10.2	26: B22.6 R11.0	29: B33.5 R9.5	31: B36.8 R6.4
18: B4.4 R15.6	23: B14.1 R20.3	27: B22.6 R23.9		
19: B4.4 R22.8	24: B14.1 R30.6			
20: B4.4 R31.9				

The sample number is followed by the sample name. The sample names are interpreted in the following manner: Sample number 19, "B4.4 R22.8", is first deesterified 4.4% with a blockwise deesterification followed by 22.8% by random deesterification. It is produced by random deesterification of sample number 16 which, in turn, is produced from the remethylated pectin. The quantities are used as the reference values for quantification.

Identification in Table 1 is based on the following convention: From the remethylated pectin (Sample number 33) the samples 1, 5, 9, 12 and 14 are produced by deesterification with a pectin esterase, known to deesterify in a predominantly random fashion. The samples are given the names "R" (for Random) followed by the degree of *de*esterification (%DDE) induced—not to be confused with the degree of esterification (%DE) defined previously and commonly used to describe pectin. All samples (1, 5, 9, 12 and 14) are produced directly from the remethylated pectin. The five samples produced are then used as starting material for a deesterification process with a block methyl esterase derived from papaya fruit, known to deesterify in a predominantly blockwise fashion. In this step samples 2–4, 6–8, 10–11, 13 and 15 are produced. The samples are named from the starting material followed by the degree of blockwise (B) deesterification. For example, sample number 10 with the name "R21.9 B9.7" is produced by deesterifying the remethylated pectin 21.9% in a random fashion (producing the sample "R21.9 B0.0"), followed by deesterifying that sample 9.7% in a blockwise fashion. The resulting degree of esterification is therefore the %DE of the remethylated pectin minus the deesterification induced, in this example 93.8% − 21.9% − 9.7% = 62.2%. The amount of deesterification induced is verified in each step by titration for all the samples, and the values are used as reference values for quantification. The samples 16–31 are produced in a likewise manner, the only difference being that they are initially subjected to a blockwise deesterification followed by a random deesterification, and hence will be named with B first and R

second. In that way all samples produced are subjected to a maximum of two deesterification steps. The range of %DE investigated here is from 93.8% down to 47.3%.

2.2. Data collection and preparation

Dissolved pectin is injected from a sample loop into a fixed flow carrier stream containing a dye that binds to pectin with high affinity. The carrier stream leads the pectin to a reactor vessel with a fixed volume, which is large relative to the carrier flow rate. The reactor is continuously stirred. The pectin is slowly diluted in the reactor as a waste stream continuously exits the reactor. Pectin is added in a surplus concentration relative to the amount of dye present in the reactor, so immediately after injection, all dye present binds to the pectin. The dye is known, from previous work, to change conformation depending on the immediate surroundings resulting in a colour change. The observed colour depends on the deesterification pattern on the pectin backbone, i.e. whether a dye-molecule binds next to other dye-molecules or not.

UV–VIS transmission spectra are continuously recorded as the pectin is diluted, and the combined spectra form a landscape. For every sample run, the raw data matrix collected consists of 333 time points and 2048 wavelengths. In order to facilitate the data analysis, the data was reduced by averaging wavelengths and time points and cropped by eliminating wavelengths with no relevant information. Time points at the start and the very end were also eliminated as they contain artefacts of the injection and mixing of the sample. The reduced data set of all the 33 samples (including the remethylated and the mother pectin) is of size 33 samples × 77 time points × 145 wavelengths. Triplicate measurements are obtained for every sample; and one of these measurements is used for model development and the two other ones are used as an initial test set.

The data was pre-processed using Matlab version 7.0.4.365 (R14 SP2) from The MathWorks and the PLS-Toolbox, version 3.5.2 from Eigenvector Research. In order to compensate for small run-to-run variations (minor baseline drift and small offsets in dye concentration), the data was corrected by fitting a straight line through all times of the first and last wavelength of the individual sample landscape and subtracting the resulting square made by connecting these time points from the landscape. The corrected data matrix was then scaled to the same intensity of wavelength number 55 (which is the peak of the largest dye peak) at the last time point where only unbound dye is observed.

Flow rates, absolute wavelengths and times, volumes and the dye involved are not given for reasons of confidentiality.

3. Methods

The landscapes of the 33 samples (using one replicate measurement per sample only) were analysed using Multivariate Curve Resolution (MCR-ALS) version 1.0.0 [21] and PARAFAC2 [22], and the results are published in the preceding article comparing the ability of the two algorithms to resolve this particular type of data. The data were resolved in common spectra for all 33 samples and either concentration profiles (MCR-ALS) or second and third mode loadings (PARAFAC2), which in turn can be converted to concentration profiles for comparison. Both algorithms succeed in describing the data well with either three or five components whereas four components provided models that were not meaningful. In this quantitative study, the three and five component MCR-ALS models with non-negativity constraints will be examined together with the unconstrained three and five component PARAFAC2 models.

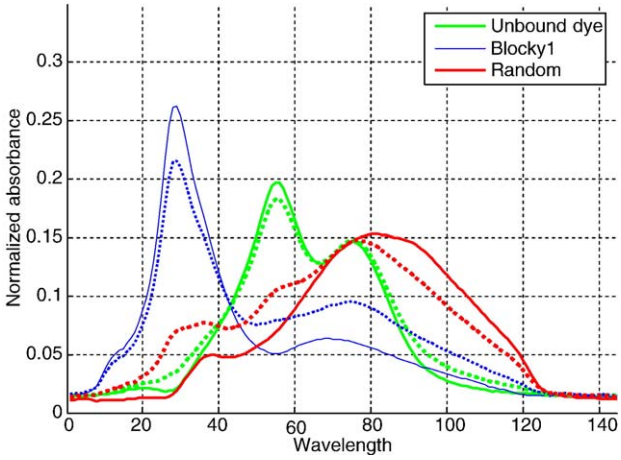


Fig. 2. Resolved spectra for the three component models. Non-negative constrained MCR-ALS (solid lines) and unconstrained PARAFAC2 (dotted lines).

In Fig. 2 the resolved spectral profiles using unconstrained PARAFAC2 or non-negativity constrained MCR-ALS from the 33 sample runs used in the calibration set are shown. Three chemical components have been resolved. The three component models have information about the unbound dye returning to the system, a component attributed to the presence of blocks (the Blocky1 profile) and a component attributed to the presence of random deesterification (the Random profile).

Even though the PARAFAC2 model is unconstrained, the algorithm converges to a solution where all three resolved spectral profiles are non-negative.

In Fig. 3, the resolved concentration profiles can be seen. The concentration profiles resolved by PARAFAC2 have negative values for considerable time stretches. This is balanced out with higher resolved concentration values of the Unbound Dye, Blocky1 and Random component compared to the profiles resolved by MCR-ALS where non-negativity constraints have been applied.

The unbound dye has a sigmoidal profile. It is close to zero at the beginning of the dilution where all the dye is supposed to be bound to the pectin. Further in the sample run, when the concentration of pectin is decreasing, the resolved concentration profile of the unbound dye increases to about two units. The time where the concentration profile of the unbound dye starts to increase is related to the total deesterification (%DDE) of the particular pectin sample. The more deesterified the pectin is, the later the profile of the unbound dye starts to increase. Three samples have a high resolved concentration profile already from the start of dilution. They are the profiles of sample number 1, 5

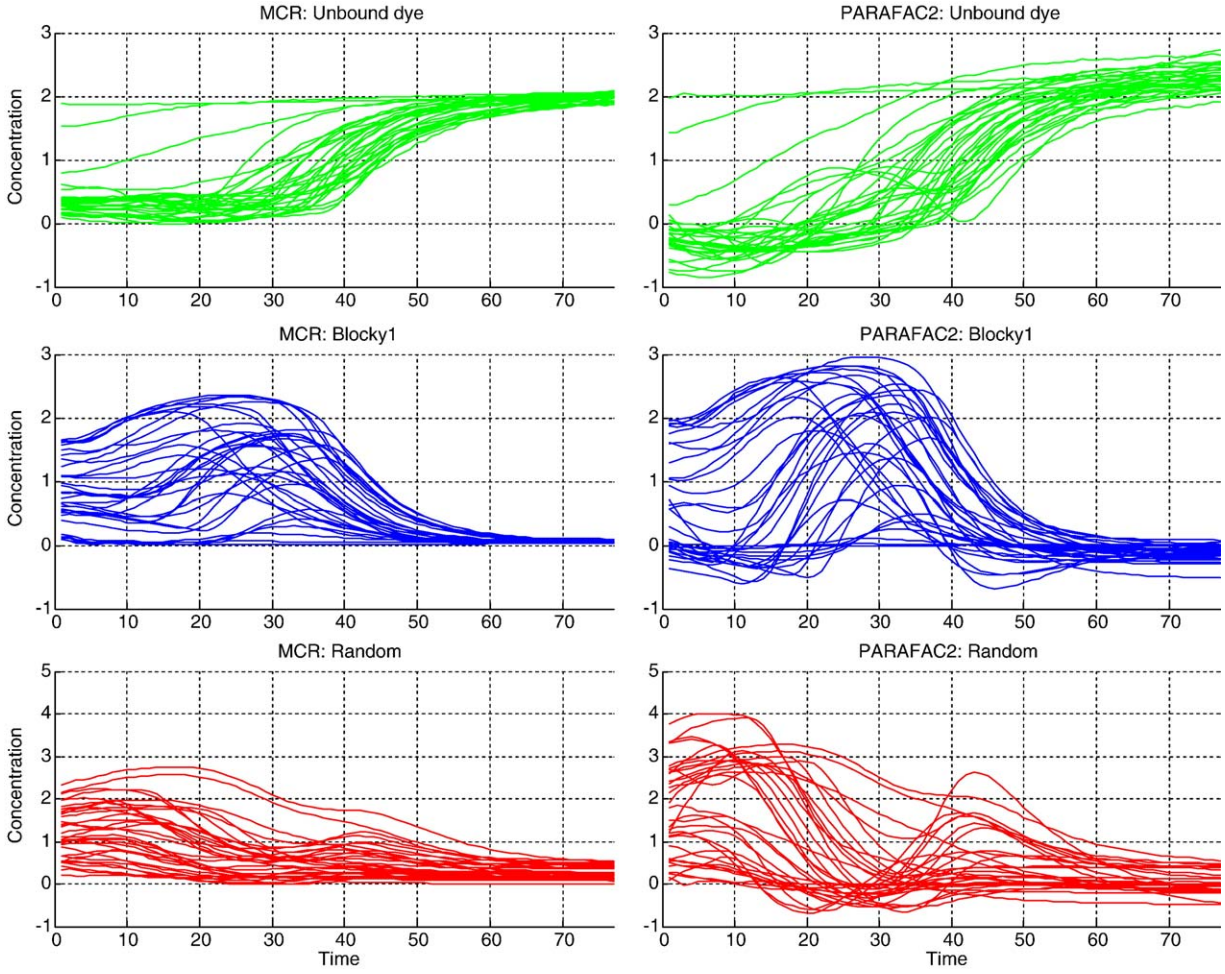


Fig. 3. Resolved concentration profiles for the three component models. MCR-ALS profiles are to the left and PARAFAC2 profiles are to the right.

and the Remethylated pectin, number 33, which are all the samples with a very low degree of deesterification.

The Blocky1 profile initially has a low or intermediate resolved concentration, after which it increases and then drops towards zero. Both the time this phenomenon starts to increase, as well as the level it increases to, is dependent on the amount of blockiness induced by deesterification. The overall relationship is not straightforward.

The random profile initially has a rather high resolved concentration. It may increase slightly before it drops. In some samples there is a tendency for the concentration to modestly increase again around time point 40. The bimodality is more pronounced in the concentration profiles resolved by PARAFAC2. This is unexpected as the concentration of pectin is decreasing exponentially due to the constant outlet flow of the reactor vessel. Examples of the data acquired, profiles from the five components unconstrained PARAFAC2 model and further interpretation of the components extracted is discussed in the preceding article [15].

As mentioned before, the entire sample set consists of 33 samples, run in triplicates. One run of each sample is used for model development as a calibration set for unfold- and multi-way PLS models as well as for MCR-ALS and PARAFAC2 modelling with subsequent PLS regression on the identified concentration profiles. The models are then applied to the two remaining sample runs, which serve as a test set. The triplicates are made from the same samples prepared by the deesterification procedure, but run on different dates, using newly prepared carriers and dye mixtures, thus testing the repeatability of the instrumental method and all modelling steps but not the sample preparation. The test of the regression models made from the spectra are straightforward, but to test the models developed on MCR-ALS or PARAFAC2 resolved concentration profiles will require that the MCR-ALS/PARAFAC2 models developed on the calibration set, are applied to the test set. For MCR-ALS this is done from the resolved spectral profiles ($\mathbf{S}_{cal}$) from the calibration set by using the following commonly used least squares estimate, where the $\mathbf{C}_{test}$ is the estimated concentration profiles and $\mathbf{D}_{test}$ is the matrix of the augmented landscapes of the test set:

$$\mathbf{C}_{test} = \mathbf{D}_{test}\mathbf{S}_{cal}\left(\mathbf{S}_{cal}^{T}\mathbf{S}_{cal}\right)^{-1} \quad (1)$$

No constraints are applied in this step, so even though constraints (e.g. non-negativity) were applied in the calibration model development and influenced the resolved spectral profiles, it is not certain that the constraints will be fulfilled for the test set. And, indeed, several samples in the test set are

resolved to have numerically small negative concentration profiles for some time points in one or more of the resolved components.

The application of the PARAFAC2 models resolved using the algorithm available at [22] is more complicated, as the test set loadings should be estimated from new landscapes keeping the spectra from the calibration set fixed. Such an algorithm is available as part of the PLS-Toolbox (version 3.5.2 from Eigenvector Research). All PARAFAC2 models fitted are unconstrained as the calibration set.

4. Results and discussion

Calibration models are made to quantify the total degree of deesterification (%DDE) and for the amount of blocky (%B) and random (%R) deesterification induced. This can be carried out using various strategies, see Fig. 4. In this article the following will be examined: 1) Unfolding the data-matrix acquired for each sample run to a vector and use Partial Least Squares (PLS) Regression [23,24] with the reference values as dependent/response variables; 2) Use Multi-way PLS [25] also called N-way PLS or simply N-PLS to correlate the landscapes of spectra directly to the reference values; 3) Reduce the acquired spectral information by unfolding only the concentration profiles identified by MCR-ALS or PARAFAC2 and use that in a PLS regression rather than the whole landscape and finally; 4) Reduce the spectral information even further by summing up or integrating the mentioned individual concentration profiles to just one number per profile identified by MCR-ALS and PARAFAC2 for the sample run, or alternatively using the identified score values (also referred to as loadings from the sample mode) from the PARAFAC2 model.

Results of the calibrations are reported in terms of the Root Mean Square Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,\mathrm{p}} - y_{i,\mathrm{r}})^2}{n}} \qquad (2)$$

where $y_{i,\mathrm{p}}$ for sample number i is the predicted value, $y_{i,\mathrm{r}}$ the actual reference value for sample number i, and n the total number of samples. The RMSE-values are in the same units and scale as the reference values. The chosen number of PLS-components for each model is identified by full cross validation of the samples in the calibration set. The estimate of the prediction error is reported as the Root Mean Square Error of Cross Validation (RMSECV). The number of PLS-components chosen does not necessarily correspond to the lowest RMSECV identified, as choosing a parsimonious model with few PLS-components is prioritized. For all models, the chosen number of PLS components falls between one and eight.

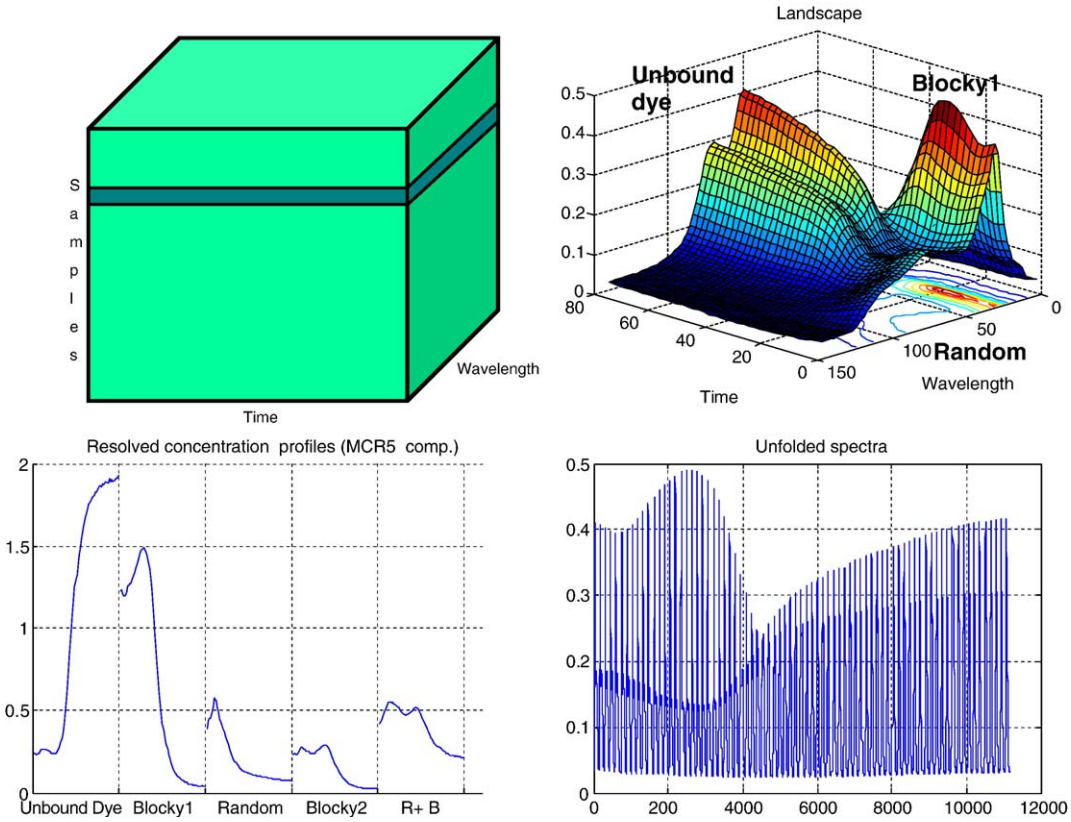


Fig. 4. Different representations of sample 6 “R13.8 B7.8”. As a slab in a data cube (top left), as a landscape on top of a contour plot with the major components labelled (top right). The same landscape is presented as an unfolded vector (bottom right) and represented as an unfolded vector of concentration profiles (bottom left), from a five component MCR-ALS solution (each composed of 77 time points) labelled according to the interpretation of the concentration profiles by its resolved spectra (not shown).

Table 2
Calibration results

Model	%DDE				%B				%R			
	#PLS-comp.	RMSEI	RMSECV	RMSEP	#PLS-comp.	RMSEI	RMSECV	RMSEP	#PLS-comp.	RMSEI	RMSECV	RMSEP
Unfold PLS on landscapes	7	2.7	3.3	2.7	4	2.5	2.7	3.5	8	2.6	3.3	3.7
N-PLS on landscapes	3	4.6	5.0	3.8	5	2.3	2.9	3.7	3	6.6	7.0	4.4
Unfold C-profiles MCR3 PLS	5	3.3	4.1	4.3	4	2.7	3.3	3.7	7	2.2	4.8	3.8
Unfold C-profiles MCR5 PLS	5	2.9	3.7	3.0	4	2.2	2.6	2.7	6	2.3	4.9	3.3
Unfold C-profiles P2-3 PLS	7	3.2	3.8	3.5	5	3.5	4.2	3.7	3	4.9	5.7	5.1
Unfold C-profiles P2-5 PLS	5	2.2	3.0	4.0	5	3.2	3.2	4.4	5	4.1	4.2	5.6
PLS[a] (areas) MCR3	6	4.7	5.0	5.4	5	6.7	6.4	6.2	2	6.4	6.9	6.2
PLS[b] (areas) MCR5	6	4.3	5.0	5.1	5	6.7	7.3	6.2	4	6.4	6.7	5.9
PLS[a] (areas) P2-3	5	5.1	5.8	6.0	5	6.2	5.5	5.8	1	6.3	7.1	6.7
PLS[b] (areas) P2-5	2	4.7	5.8	6.7	5	7.4	5.5	8.2	5	5.7	6.9	6.0
PLS[a] (scores) P2-3	2	8.4	9.2	20.5	3	6.3	8.5	14.2	3	5.5	6.7	11.0
PLS[b] (scores) P2-5	4	6.7	7.8	17.3	4	4.5	5.5	12.3	6	6.6	6.8	19.4

MCR3, P2-3, MCR5 and P2-5 refers to non-negativity constrained MCR-ALS and unconstrained PARAFAC2 models with respectively 3 or 5 resolved concentration profiles. See the text for a full explanation of the modelling strategy and the RMSE values tabulated.

[a] The PLS models using the areas and score values are expanded with all quadratic and interaction terms of combinations of the resolved profiles.

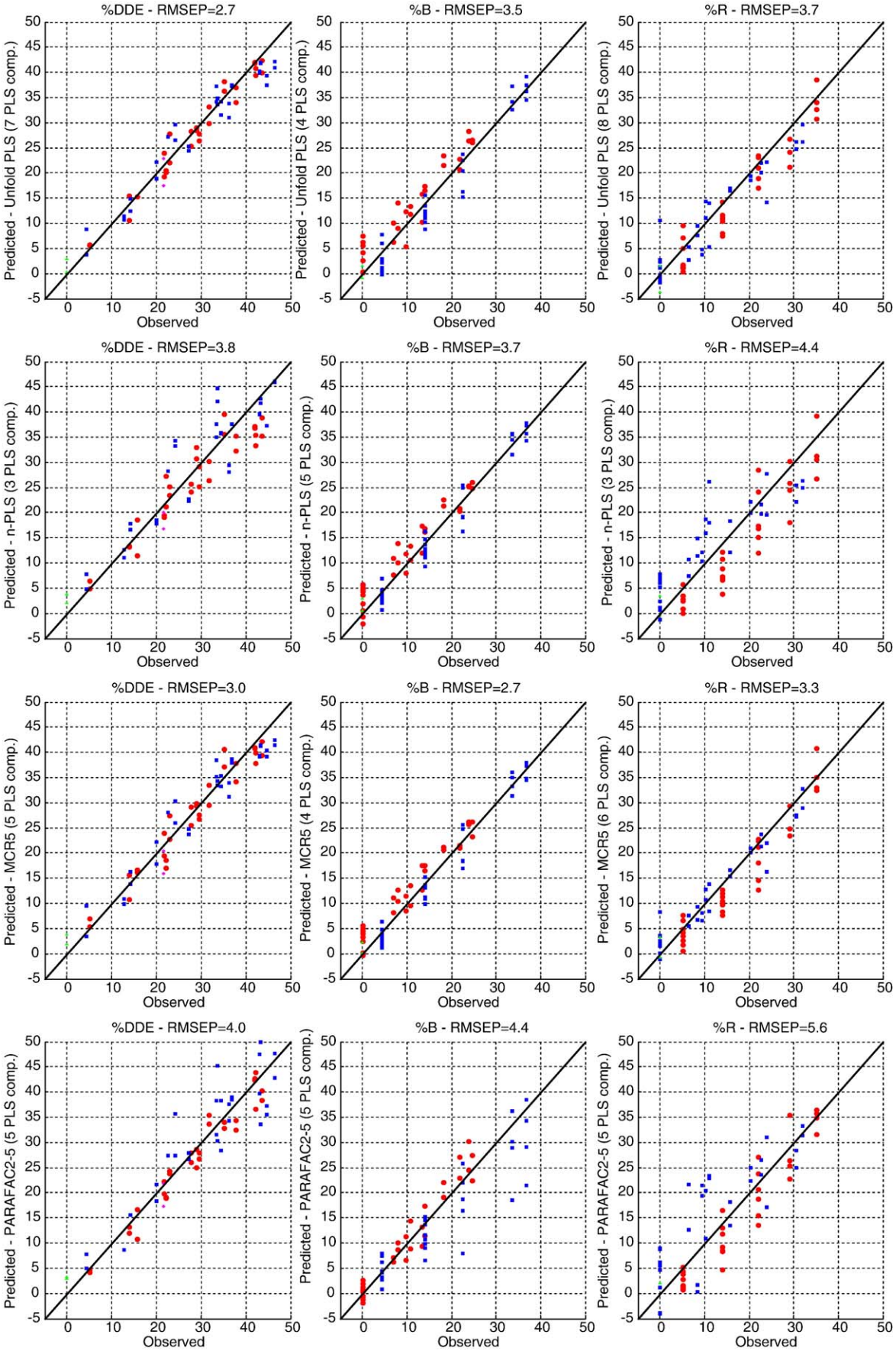
[b] The PLS models using the areas and score values are expanded with the quadratic terms of the resolved profiles.

All models are tested in two ways. First, 10 samples are removed from the calibration set (sample number 2, 4, 5, 7, 13, 17, 20, 21, 23 and 29), attempting to represent all facets of the entire sample set, and used as an "internal" test set. Models are developed on the remaining 23 samples (including the remethylated and mother pectin—there are however no reference %B and %R values for the mother pectin thus invalidating this sample for calibration of those parameters) and tested on the internal test set. The same number of PLS-components chosen by the full cross validation of the entire calibration set is used. The RMSE error term is reported as the Root Mean Square Error of Internal test set (RMSEI). In this way of validating the models, the replicates of the same pectins are not used for validation purposes, but this "internal" test set is investigating the adequacy of the choice of number of components in the PLS modelling step only, as no intermediate recalculations of PARAFAC2 and MCR-ALS models are done based on only the reduced calibration set.

In the second test, the calibration models are applied to the two remaining samples from the triplicate runs, not used in any modeling step. This gives 2*33 samples (still including the remethylated and mother pectin) except for one outlier that was removed. The RMSE error term is reported as the Root Mean Square Error of Prediction (RMSEP). This parameter tests the instrumental method (the triplicates are run on different days, using different carrier preparations and two operators have been involved), the resolution by PARAFAC2 and MCR-ALS and the PLS modelling step. The samples are identical however, and the reference values used are the same (an average of triplicate measurements of the reference titration method). The same number of PLS-components chosen by the full cross validation of the entire calibration set is used for the prediction test set. The result of all the calibrations can be seen in Table 2 and in the predicted versus measured plots for selected models in Fig. 5.

The RMSE values should be seen in relation to the ranges spanned by the reference values that are 46.5%, 36.8% and 35.0% for %DDE, %B and %R respectively. Almost consistently RMSEI values are lower than the RMSECV (0.5% on average). This is a somewhat unexpected result as the RMSECV values are predicted from a model built on 31 or 32 samples leaving only the cross-validated sample out. The RMSEI values are predicted from a model built on only 22 or 23 samples. But the 10 samples constituting the internal test set happen to be predicted better than the average of all samples. The stratified selection of the internal test set does not include many corners in the design, rather the interior of the designs are represented, which predicts very well and this is likely the reason for the lower RMSEI values. The RMSEP values are in general of the same size as the RMSECV with some noteworthy exceptions. This indicates that the instrumental method and all modelling steps work well and do not lead to overfit.

The unfold PLS models have in general the lowest RMSE-values, on average 3.0% but also requires the highest number of PLS-components to model the data. The landscapes are unfolded from 77 times 145 wavelengths to spectra with 11,165 variables, which is almost 30–1200 times more variables than the other models. Clearly this increases the chance of overfit, but from the table is seen not to happen presumably because the variation in the data is highly relevant. For calibration purposes the way the landscapes are unfolded is immaterial. No further pre-processing of the spectra and variable selection is done in order to compare the models on more similar conditions. N-PLS uses fewer PLS-components but does yield models with worse RMSE-values. In particular the prediction of %DDE and %R is poor, whereas the prediction


%DDE - RMSEP=2.7
Predicted - Unfold PLS (7 PLS comp.)
%B - RMSEP=3.5
Predicted - Unfold PLS (4 PLS comp.)
%R - RMSEP=3.7
Predicted - Unfold PLS (8 PLS comp.)
%DDE - RMSEP=3.8
Predicted - n-PLS (3 PLS comp.)
%B - RMSEP=3.7
Predicted - n-PLS (5 PLS comp.)
%R - RMSEP=4.4
Predicted - n-PLS (3 PLS comp.)
%DDE - RMSEP=3.0
Predicted - MCR5 (5 PLS comp.)
%B - RMSEP=2.7
Predicted - MCR5 (4 PLS comp.)
%R - RMSEP=3.3
Predicted - MCR5 (6 PLS comp.)
%DDE - RMSEP=4.0
Predicted - PARAFAC2-5 (5 PLS comp.)
%B - RMSEP=4.4
Predicted - PARAFAC2-5 (5 PLS comp.)
%R - RMSEP=5.6
Predicted - PARAFAC2-5 (5 PLS comp.)
Observed

of %B is comparable to the performance of other models. The prediction of the test set from the landscapes can be seen in the top two rows of Fig. 5.

Predictions based on the unfolded concentration profiles resolved by MCR-ALS and PARAFAC2 are good. MCR-ALS models have on average 0.7% lower RMSE-values than PARAFAC2 (an improvement from on average 4.0% to 3.3%), and the models based on five resolved concentration profiles give on average 0.5% lower RMSE-values (an improvement from on average 3.9% to 3.4%). The prediction performance of five concentration profiles resolved by MCR-ALS is comparable to the models based on the unfolded landscapes (3.1% compared to 3.0%). The unfolded concentration profiles have 5*77=385 variables compared to the 11165 variables of the unfolded landscapes. The prediction of the test set from the PLS models based on the 5 concentration profiles resolved by PARAFAC2 and MCR-ALS can be seen in the bottom two rows of Fig. 5.

When the information in the resolved concentration profiles are reduced even further to just one number per concentration profile equivalent to the individual sum of the concentration profile or the area under the concentration profile curve, predictions do not go well as the spectral response seen in the landscapes do not have a linear relationship with regard to the type of deesterification induced. The direct correlation coefficient between the areas of the different concentration profiles seen in e.g. Fig. 3, and the induced deesterification are ranging between 0.55 and 0.82. For instance the area of the concentration profile identified as the unbound dye resolved by a 3 component PARAFAC2 model correlates with −0.816 with the %DDE, which verifies that the more deesterified the sample is, the later the concentration profile identified as the unbound dye appears, the smaller the area will be when summing up on that particular concentration profile. Hence, the correlation is expected to be negative. The numerically highest correlations are seen with the %DDE and the lowest correlations are seen with %B, which concentration profile is very non-linear in intensity. In order to compensate for the inherent nonlinearity the PLS-models based on the areas or score values only are expanded with their interaction terms (3 profile models only) and quadratic terms (3 and 5 profile models) before modelling the relationship to %DDE, %B and %R with PLS. In this way the 3 profile models will be modelled from a total of nine PLS variables which are the three area/score values by themselves, the three quadratic terms, and the three possible combinations of the area/score values multiplied. The 5 profile models will be modelled from a total of 10 variables, which are the five area/score values and their five quadratic terms. This improves the predictions, but the results are not satisfying and the average RMSE-values are still around 6.0%.

Score values from PARAFAC2 and multivariate models in general are expected to correlate to reference values, but this is not found to be the case in this example. Prediction errors are as bad as 11–20% for the test set. In this case the time shifts in the concentration profiles, which are well modelled and useful for curve resolution by MCR-ALS and PARAFAC2, turn out to be very important for this application. It is not a property that should be "corrected for" and eliminated in the case of quantification. Because PARAFAC2 allows for independent time profiles, the variation is kept in the time profiles and is not reflected in the PARAFAC2 score values. Therefore the computed concentration profiles from PARAFAC2 works well for prediction but the score values do not.

When looking at the reference values %DDE and %B is predicted well whereas the prediction of %R is more difficult which is reflected in higher RMSE-values. The random spectral profiles resolved is a considerably wider peak than the peak of the dye and blocky profile (see Fig. 2), so it could be that the dye bound to the random sites on the pectin backbone is less well defined. Closer inspection of Fig. 5 also reveals that samples first deesterified randomly followed by a blocky deesterification (the Random first samples) seems to be better predicted than the Blocky first samples.

Inspection and interpretation of the PLS score-plots can provide important information about the relation between the samples and the reference values and overall validity of the models. In Fig. 6 score plots from the models to %DDE from the landscapes directly (unfolded or by n-PLS) and three profiles MCR-ALS and PARAFAC2 models are shown. All four score plots are similar to each other and directly reflect the design seen in Table 1. The samples in the score plots are connected with lines according to the layout of the design. PLS-component #1 reflects the information about %R and PLS-component #2 (#3 for n-PLS) reflect information about the %B induced. Since the PLS-model is made to the total degree of deesterification (%DDE) only, no information about the two ways of deesterification influence the decomposition of the variables. The fact that the score plot of the unfolded concentration profiles of the MCR-ALS and PARAFAC2 models are so similar to the models based on the landscapes is also evidence that the MCR-ALS and PARAFAC2 models are successful in modelling the landscapes in terms of fit and relevance for the reference values.

Also possible to assess from score plots is the replicate variation. In Fig. 7, the score plot of PLS component #1 and #2 of samples from the test set from a PLS model to %DDE based on three unfolded MCR-ALS concentration profiles resolved are shown. It can be seen that the replicate variation is relatively large compared to the overall variation. In this score plot (accounting for 84% of the variation of the concentration profiles and 83% of the variation in %DDE) there is no systematic variation between the replicates, as there is no uniform direction or trend of the lines on the figure. The variation is hence a property of the instrumental method, the

Fig. 5. Predicted versus measured reference values for models based on the landscapes directly using unfold PLS or n-PLS (top two rows) and using MCR-ALS and PARAFAC2 to resolve five concentration profiles which are unfolded before using PLS (bottom two rows). The samples are marked according to the deesterification sequence (▲ Remethylated pectin, ◆ Mother pectin, ● Random first, ■ Blocky first).

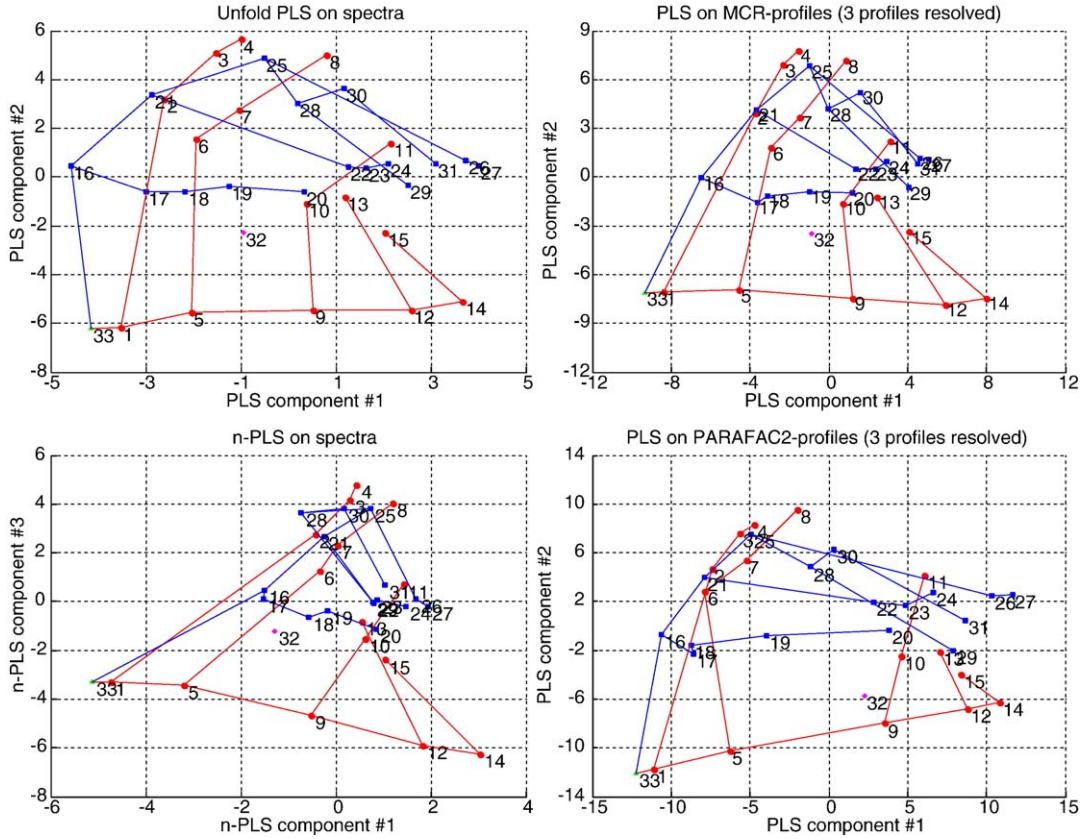


Fig. 6. Score plots from four PLS models of %DDE. All score plots reflect the design presented in Table 1. As the models are calibrated with the %DDE induced, no information about the deesterification sequence is imposed on the models by the reference value. The samples are marked according to the deesterification sequence (▲ Remethylated pectin, ♦ Mother pectin, ● Random first, ■ Blocky first) and numbered according to Table 1.

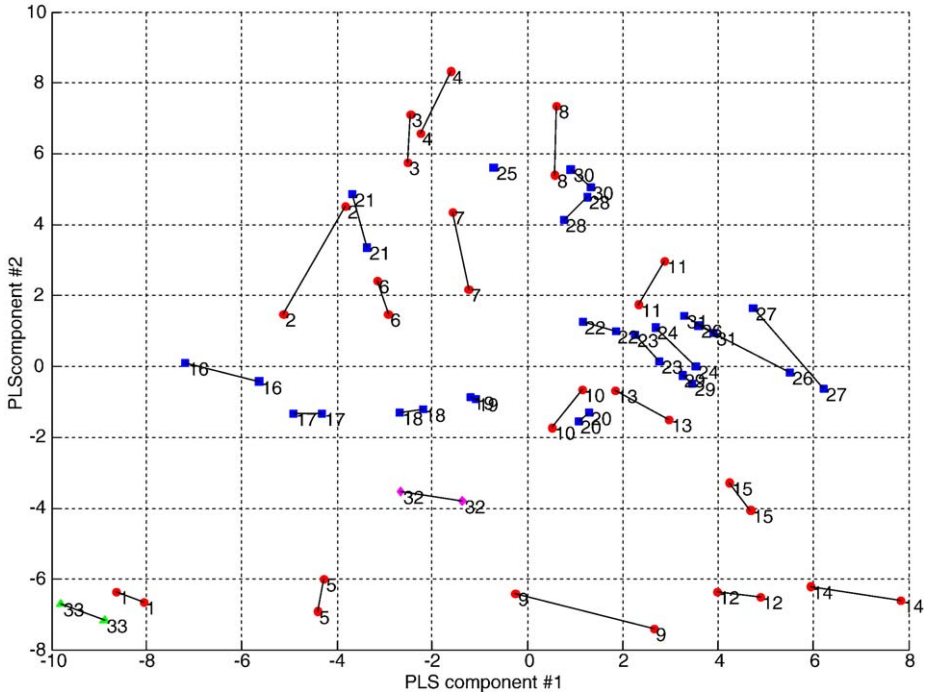


Fig. 7. Score plot of PLS component #1 and #2 of samples from the test set predicted with a PLS model of %DDE based on three unfolded MCR-ALS resolved concentration profiles. The lines indicate distances between two runs used as test set from the triplicate measurement of the same samples and give an idea about the repeatability of the instrumental method and modelling step. The samples are marked according to the deesterification sequence (▲ Remethylated pectin, ♦ Mother pectin, ● Random first, ■ Blocky first) and numbered according to Table 1.

MCR-ALS estimation of the concentration profiles and the PLS modelling step, extracting the variation in the concentration profiles, which describe the variation in %DDE.

The variation seen in the concentration profiles translates into an average of 3.0 %DDE between the predicted two independent runs constituting the test set, which is rather large compared to the overall RMSEP value of the model of 4.3 %DDE. A Principal Component Analysis (PCA) of the unfolded spectra would focus on the contribution of the instrumental method by itself. Also a raw data inspection reveals that there is a noticeable variation of the raw data, so improvements in the instrumental method will benefit the quantification.

5. Conclusion

PLS has been applied successfully to the dilution profiles of 33 pectin samples injected in triplicates quantitatively in a system with a constant concentration of dye, binding site-specifically to the galacturonic acid group of the pectin carbohydrate backbone. Variations in the pectin samples induced by enzymatically deesterifying in different deesterification patterns (blocky or random) lead to different conformations and dilution profiles of the dye which are detected by UV–VIS spectroscopy. This gives the data a three-way structure of sample × time × wavelength. Results from titration of the pectin give the degree of deesterification and thereby the effect of the enzymatic deesterification induced, which is used as reference values as either degree of deesterification (%DDE), degree of blocky (%B) or degree of random (%R) deesterification induced. The standard deviation of the reference titration is 0.7%.

Both the PARAFAC2 and the MCR-ALS algorithms successfully resolve three concentration profiles, which can tentatively be attributed to three phenomena, namely a fingerprint of blocky acid groups, random acid groups and the unbound dye, which is seen when the concentration of pectin is diluted to a sufficiently low concentration. Further two minor components to be resolved by MCR-ALS describing minor spectral wavelength shifts and/or artefacts observed at very low pectin concentrations when the unbound dye starts to increase in concentration. The additional components cannot be resolved very well by PARAFAC2 where signs of overfit are observed. The extra step of modelling the landscapes by MCR-ALS or PARAFAC2 provides an in-depth understanding of the recorded landscapes and offer possibilities of curve resolution and interpretation not clearly seen from the landscapes or PLS/N-PLS parameters. The best predictive performance is reached using PLS on the unfolded landscapes directly where %DDE, %B and %R in the test set made of two of the three triplicates have RMSEP values of 2.7%, 3.5% and 3.7%, respectively. Almost as good predictive ability has PLS on the unfolded MCR concentration profiles, which performs better than PLS-models on the unfolded PARAFAC2 profiles. Five resolved and unfolded concentration profiles predict better than three resolved and unfolded concentration profiles when resolved by MCR. When resolved by PARAFAC2 the prediction errors are higher using five profiles compared to three. N-PLS does not perform as well as the unfolded MCR concentration profiles with RMSEP values of 3.8%, 3.7% and 4.4% in comparison, but these values are still better than the unfolded concentration profiles resolved by PARAFAC2 with the exception of the %DDE value predicted with three unfolded PARAFAC2 profiles. When the resolved concentration profiles are reduced even further to just a number by taking the simple sum of the profile, the predictive ability deteriorates.

It can therefore be concluded that the landscapes contain relevant information for the determination of %DDE, %B and %R, and that (qualitative) information can be successfully extracted using MCR-ALS and PARAFAC2, preserving (quantitative) concentration profiles which perform nearly as well using only a fraction of the variables. The resolved concentration profiles from MCR-ALS perform better than PARAFAC2, which could be due to inherent higher flexibility in the MCR-ALS algorithm, which resolves concentration profiles with less restrictions compared to PARAFAC2. The RMSEP errors of 3–4% as presented are still high compared to the standard deviation of 0.7% of the reference titration method for determining the degree of esterification. The reference method is however unable to measure the distribution of the ester groups. A method for this has been proposed using high resolution ^{1}H NMR [6], and analysis on the same sample set partly verifies the results presented in this article.

Acknowledgements

The authors would like to gratefully acknowledge the grant made available by the Ministry of Science, Technology and Innovation to Christian B. Zachariassen, to partly cover the expenses in connection with participation in the Industrial PhD project, which is a joint effort between CP Kelco and the Quality and Technology Group, The Royal Veterinary and Agricultural University of Denmark.

At CP Kelco, many colleagues have contributed in taking this application from idea to reality. Special credit is given to the laboratory personnel, viz. Anette Møller, Britta Pedersen and Hanne Thulstrup Jensen.

References

[1] C. Rolin, in: G.B. Saymour, J.P. Knox (Eds.), Pectins and their Manipulation, Blackwell Publishing, Oxford, Great Britain, 2002, pp. 222–241.
[2] Compendium of Food Additives Specifications/add. 9, FAO Food and Nutrition Paper no. 52, Rome, Italy, 2001.
[3] C. Rolin, J. De Vries, in: P. Harris (Ed.), Food Gels, Elsevier, Amsterdam, The Netherlands, 1990, pp. 401–434.
[4] P.J.H. Daas, A.G.J. Voragen, H.A. Schols, Investigation of the galacturonic acid distribution of pectin with enzymes part 2—characterization of non-esterfied galacturonic acid sequences in pectin with endopolygalacturonase, Carbohydrate Research 326 (2000) 120–129.
[5] G. Limberg, R. Korner, H.C. Buchholt, T.M.I.E. Christensen, P. Roepstorff, J.D. Mikkelsen, Analysis of pectin structure part 3—quantification of the amount of galacturonic acid residues in blocksequences in pectin homogalacturonan by enzymatic fingerprinting with exo- and endo-polygalacturonase II from *Aspergillus niger*, Carbohydrate Research 327 (2000) 321–332.

[6] H. Winning, N. Viereck, L. Nørgaard, J. Larsen, S.B. Engelsen, Quantification of the degree of blockiness in pectins using ^{1}H NMR spectroscopy and chemometrics, Food Hydrocolloids (in press).
[7] J. Solms, H. Deuel, Über den Mechanismus der enzymatischen Verseifung von Pektinstoffen, Helvetica Chimica Acta 38 (1955) 321–329.
[8] L. Miller, J.D. MacMillan, Purification and pattern of action of pectinesterase fom *Fusarium oxysporum* f. sp. *vasinfectum*, Biochemistry 10 (1971) 570–576.
[9] R. Kohn, I. Furda, Z. Kopec, Distribution of free carboxyl groups in pectin molecule after treatment with pectin esterase, Collection of Czechoslovak Chemical Communications 33 (1968) 264–269.
[10] R.M. McCready, C.G. Seegmiller, Action of pectic enzymes on oligogalacturonic acids and some of their derivatives, Archives of Biochemistry and Biophysics 50 (1954) 440–450.
[11] S. Ishii, K. Kiho, S. Sugiyama, H. Sugimoto, Low-methoxyl pectin prepared by pectinesterase from *Aspergillus japonicus*, Journal of Food Science 44 (1979) 611–614.
[12] E.J. Lourenco, A.T. Catutani, Purification and properties of pectinesterase from papaya, Journal of the Science of Food and Agriculture 35 (1984) 1120–1127.
[13] K.S. Booksh, B.R. Kowalski, Theory of analytical chemistry, Analytical Chemistry 66 (1994) 782A–791A.
[14] R. Tauler, Multivariate Curve Resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems 30 (1995) 133–146.
[15] C.B. Zachariassen, J. Larsen, F. van den Berg, R. Bro, A. de Juan, R. Tauler, Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system, Chemometrics and Intelligent Laboratory Systems 83 (2006) 13–25.
[16] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2—Part I. A direct fitting algorithm for the PARAFAC2 model, Journal of Chemometrics 13 (1999) 275–294.
[17] R. Bro, C. Andersson, H.A.L. Kiers, PARAFAC2—Part II. Modeling chromatographic data with retention time shifts, Journal of Chemometrics 13 (1999) 295–309.
[18] R. Tauler, A.K. Smilde, B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in Multivariate Curve Resolution, Journal of Chemometrics 9 (1995) 31–58.
[19] A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures Revisiting latest trends in multivariate resolution, Analytica Chimica Acta 500 (2003) 195–210.
[20] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for Multivariate Curve Resolution in MATLAB, Chemometrics and Intelligent Laboratory Systems 76 (2005) 101–110.
[21] http://www.ub.es/gesq/mcr/mcr.htm as of June 2005.
[22] http://www.models.kvl.dk/source/ as of June 2005.
[23] M. Sjöström, S. Wold, W. Lindberg, J.A. Persson, H. Martens, A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables, Analytica Chimica Acta 150 (1983) 61–70.
[24] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130.
[25] R. Bro, Multiway calibration. Multilinear PLS, Journal of Chemometrics 10 (1996) 47–61.