Fast quality assessment of barley and wheat:

Chemometric exploration of instrumental data with single seed applications

Ph.D. thesis by

Jesper Pram Nielsen, M.Sc. Agricultural Science

Supervisor

Professor Lars Munck
Food Technology
Department of Dairy and Food Science
The Royal Veterinary and Agricultural University
Rolighedsvej 30, DK-1958 Frederiksberg C.

Preface

This thesis has been submitted as partial fulfilment of the requirements for a Ph.D. degree at The Royal Veterinary and Agricultural University, Copenhagen, Denmark. The supervisor has been Professor Lars Munck to whom I am grateful for giving me the opportunity to do this work, and for enthusiastic discussions on the "spirit of exploratory data analysis". I am greatly indebted to Lars Nørgaard for introducing me to the world of chemometrics and for bridging the interface between theory and practice in exploratory data analysis.

Bo Löfqvist, BoMill AB, Sweden is greatly appreciated for his invaluable support and encouragement. Jørgen Larsen, Carlsberg Research Laboratory is thanked for his assistance and for introducing me to the importance of malting barley quality. I am also grateful to Dominique Bertrand for inspiring co-operation during my stay at ENITIAA/INRA, Nantes, France.

The entire Food Technology group is thanked for creating a superb scientific and social environment. Dorthe, Birthe, Rasmus, Harald, Elisabeth, Frans and Gilda are especially acknowledged for their support and help, and Kirsten and Karen (former employees) are acknowledged for their technical assistance in the tedious development of the single seed laboratory analyses.

Most of the work presented in this thesis originates from the research projects: "Fast analysis methods for single seeds" (The Danish Cereal Network-B4) oriented towards development of a multifunctional analytical instrument, and "Cascade Refining of European Wheat for Production of High-Quality Products for the Paper Industry" (EU-FAIR CT96-1105) oriented towards the cereal processing and utilisation industries. The participants of these projects are warmly acknowledged for inspiring discussions and meetings, and The Directorate for Food, Fisheries and Agri Business as well as the European Commission are acknowledged for their financial support of these two projects.

Copenhagen, October 2002

Jesper Pram Nielsen

Summary

This thesis describes the development and demonstrates the potential of fast multivariate instrumental methods combined with chemometrics for monitoring, handling and utilisation of quality variations in barley and wheat. The thesis consists of seven papers (basis of the thesis) and three additional papers (appendices). In the first part of the thesis, the applied instrumental and chemometric data evaluation methods are briefly described as a background to the research papers. The new screening methods presented in the included papers are discussed in relation to their potential application within plant breeding, handling and sorting and cereal processing, all of which may assist in improving the end product quality.

One of the main aspects of this thesis is the development and application of single seed analysis. The fast instrumental methods applied, including kernel morphology by image analysis, hardness analysis and near infrared spectroscopy, have therefore been chosen, since they can provide single seed data. These fast instrumental methods are utilised as physical and chemical fingerprints which in combination with chemometrics are used for the characterisation of wheat and barley quality.

Paper 1 demonstrates an exploratory evaluation of classical analytical micro-malting data in which Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) are used for data reduction and defining of underlying functional factors. Three papers address the issue of developing screening methods for prediction of classical malting quality analyses. The use of single seed morphology and hardness data on bulk samples (Paper 2) and near infrared spectroscopy on bulk samples (Paper 3) shows potential predictive ability for the physical and chemical parts of a malt quality profile, but was not able to predict the biochemical part related to germination capacity. Paper 4 demonstrates the use of fuzzy logic for the translation of a multivariate malt quality profile into a valid overall quality index. This index proved to be predictable by near infrared spectroscopy, and thus may reduce the work devoted to quality evaluation in malting barley breeding. Paper 5 is an example of how spectroscopy and chemometrics can be used on the phenotype level to detect the expression of a regulator gene in barley.

The topic of Paper 6 is the development of non-destructive screening methods for single seed protein, vitreousness, density and hardness in wheat. Near Infrared

transmittance (NIT) spectroscopy showed excellent ability to predict single seed protein content, while the non-destructive prediction of vitreousness, density and hardness were more difficult. The poor NIT prediction of hardness, however, proved to be mainly due to inaccuracy in the single seed hardness measurements.

Finally, a chemometric multi-way study for an improved use of near infrared spectroscopy in process monitoring in wheat milling is described in Paper 7.

In addition to the seven papers, three appendices are included, as these are discussed in the thesis. In these appendices, new chemometric preprocessing methods in terms of variable selection (Appendix 1, used in Paper 3) and scatter correction (Appendices 2 and 3) are proposed. All of these proved to considerably improve predictive models based on near infrared spectroscopy.

Resumé

Denne afhandling beskriver udvikling af multivariate hurtigmetoder i kombination med kemometri og demonstrerer potentialet af disse inden for monitorering og udnyttelse af kvalitetsvariationer i byg og hvede. Afhandlingen indeholder syv artikler som er kernen, med tre associerede artikler som appendiks. Indledningsvis beskriver afhandlingen de anvendte instrumentelle analysemetoder og den tilknyttede kemometriske databehandling som baggrund for artiklerne. Dernæst diskuteres de nyudviklede screeningsmetoder i relation til deres potentielle brug inden for kornforædling, kornhåndtering, kornsortering og monitorering under korn-forarbejdning for derigennem at forbedre kvaliteten af slutprodukterne.

Et vigtigt element i dette arbejde er udvikling og anvendelse af enkeltkerne-analyser. De anvendte hurtigmetoder er derfor valgt, da de giver mulighed for at måle på enkelte kerner, enten i en population eller som reelle enkeltkerneanalyser. Metoderne er kernemorfologi målt med billedanalyse, hårdhedsanalyse og nærinfrarød spektroskopi. Disse hurtigmetoder anvendes som et fysisk/kemisk fingeraftryk af prøven, som i kombination med kemometri anvendes til at karakterisere byg- og hvedekvalitet.

Artikel 1 beskriver anvendelsen af eksplorativ dataanalyse i form af Principal Component Analysis og Partial Least Squares Regression til analyse af mikromaltningsdata, hvorved underliggende funktionelle egenskaber kan defineres. Tre artikler omhandler udvikling af hurtigmetoder til bestemmelse af maltkvalitet. Både anvendelse af bygkernemorfologi og hårdhed (Artikel 2) og nærinfrarød spektroskopi (Artikel 3) på bulk prøver kan prædiktere de fysiske og kemiske maltkvalitets-parametre, men var ikke i stand til at prædiktere de biokemiske parametre relateret til spiringskapacitet. Artikel 4 demonstrerer brugen af fuzzy logic til oversættelse af en multivariat maltkvalitetsprofil til et meningsfyldt overordnet kvalitetsindeks. Dette indeks viste sig ydermere at kunne prædikteres med nærinfrarød spektroskopi, og derved kan analysearbejdet i maltbyg-forædlingen reduceres. Artikel 5 er et eksempel på hvordan spektroskopi og kemometri kan anvendes til at detektere den fænotypiske expression af et regulator gen i byg.

Artikel 6 omhandler udvikling af ikke-destruktive screeningsanalyser for enkeltkerne protein, vitrositet, densitet og hårdhed i hvede. Nær infrarød transmittans (NIT) viste sig at være særdeles velegnet til bestemmelse af

enkeltkerne protein, mens ikke-destruktiv prædiktion af vitrositet, densitet og hårdhed viste sig mere vanskeligt. Den dårlige prædiktion af hårdhed viste sig dog primært at skyldes for stor usikkerhed på hårdhedsmålingen af enkeltkerner.

Endeligt er Artikel 7 et studie af en kemometrisk multivejsmetode til forbedret brug at nærinfrarød spektroskopi som overvågningsredskab under industriel formaling af hvede.

Ud over de syv artikler indeholder afhandlingen tre artikler, der bliver diskuteret og som derfor er gengivet i appendiks. Disse appendikser foreslår nye kemometriske forbehandlingsmetoder, nemlig variabel selektion (Appendiks 1, brugt i Artikel 3) og signal korrektion af spektrale data (Appendiks 2 og 3). Alle disse metoder viste sig at give betragtelige forbedringer af prædiktionsmodeller baseret på nærinfrarøde spektre.

List of publications included in the thesis

This thesis is based on the publications listed below. These publications are included in full text and will be referred to as Papers 1 through 7 throughout the text.

Paper 1:

Jesper Pram Nielsen and Lars Munck. Evaluation of malting barley quality using exploratory data analysis. I. Extraction of information from micro-malting data of spring and winter barley. *Journal of Cereal Science*, submitted

Paper 2:

Jesper Pram Nielsen. Evaluation of malting barley quality using exploratory data analysis. II. The use of kernel hardness and image analysis as screening methods. *Journal of Cereal Science*, submitted

Paper 3:

Jesper Pram Nielsen and Lars Munck. Prediction of malt quality on whole grain and ground malt using near infrared spectroscopy and chemometrics. In Near Infrared Spectroscopy: Proceedings of the 9th International Conference, Ed. by A.M.C. Davies and R. Giangiacomo. NIR Publications, Chichester, UK, pp. 709-713 (2000)

Paper 4:

Jesper Pram Nielsen, Rasmus Bro, Jørgen Larsen and Lars Munck. Application of fuzzy logic and near infrared spectroscopy for malt quality evaluation. *Journal of the Institute of Brewing*, submitted.

Paper 5:

Lars Munck, Jesper Pram Nielsen, Birthe Møller, Susanne Jacobsen, Ib Søndergaard, Søren B. Engelsen, Lars Nørgaard and Rasmus Bro. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Analytica Chimica Acta*, 446, 171-186 (2001).

Paper 6:

Jesper Pram Nielsen, Dorthe Kjær Pedersen and Lars Munck. Development of non-destructive screening methods for single kernel characterisation of wheat. *Cereal Chemistry*, accepted.

Paper 7:

Jesper Pram Nielsen, Dominique Bertrand, Elisabeth Micklander, Phillip Courcoux and Lars Munck. Study of NIR spectra, particle size distributions and chemical parameters of wheat flours: a multi-way approach. *Journal of Near Infrared Spectroscopy*, **9**, 275-285 (2001).

List of additional publications

The following publications are not directly included in the thesis. Since they are relevant and discussed in the thesis, they are added as appendices.

Appendix 1:

Lars Nørgaard, Arild Saudland, Jesper Wagner, Jesper Pram Nielsen, Lars Munck and Søren Balling Engelsen. Interval partial least squares regression (iPLS): A comparative chemometric study with an example from near infrared spectroscopy. *Applied Spectroscopy* 50 (4) 413 - 419 (2000).

Appendix 2:

Harald Martens, Jesper Pram Nielsen and Søren Balling Engelsen. (2002). Light scattering and light absorbance separated by Extended Multiplicative Signal Correction (EMSC). Application to NIT analysis of powder mixtures. *Analytical Chemistry*, accepted.

Appendix 3:

Dorthe Kjær Pedersen, Harald Martens, Jesper Pram Nielsen and Søren Balling Engelsen. (2002). Near infrared absorption and scattering separated by Extended Inverted Signal Correction (EISC). Analysis of NIT spectra of single wheat seeds. *Applied Spectroscopy*, 56 (9) 1206-1214.

Table of contents

Abbreviation list

AACC	American Association of Cereal Chemists
COMDIM	Analysis of Common Dimensions and Specific Weights
EBC	European Brewery Convention
HI	Hardness Index (SKCS)
iPLS	Interval Partial Least Squares (Regression)
MSC	Multiplicative Signal Correction
NIR	Near Infrared Reflectance
NIT	Near Infrared Transmittance
PC	Principal Component
PCA	Principal Component Analysis
PLSR	Partial Least Squares Regression
RE	Relative Error
RHI	Relative Hardness Index (SKCS on barley)
RMSECV	Root Mean Squared Error of Cross Validation
RMSEP	Root Mean Squared Error of Prediction
siPLS	Synergy Interval Partial Least Squares (Regression)
SKCS	Single Kernel Characterization System

1. Introduction

At present, cereals in industry are processed to detailed quality specifications which demand a range of analyses on the raw material, the process intermediates and on the final products. Three main aspects are of importance in the determination of the usability and value of grain raw material, namely its composition, its functionality and its safety. Grain composition is important from a nutritional point-of-view and, to a certain extent, to functionality. Grain composition includes parameters such as moisture, protein, oil, fibre, starch or other carbohydrates. Functionality means the capability of the grains to serve as raw material in food and feed processes. Grain functionality is ideally evaluated by laboratory-scale preparation of the desired end product such as micro-malting and mashing or test-baking. Kernel hardness, ß-glucan content, enzymatic activity and germination rate highly influence the functionality of barley in malting, while kernel hardness, gluten quality and gluten strength influence the functionality of wheat for milling and baking. The safety aspect includes determination of the presence of toxic substances, such as mycotoxins, the presence of insects and the presence of residues of pesticides. The safety aspect will not be dealt with in this thesis.

Grain quality thus includes physical, chemical and functional aspects depending on the intended purpose. A range of parameters are involved in a full quality characterisation of cereals (see the manuals of the AACC (Anonymous, 1983) and EBC Analytica (Anonymous, 1987)). These traditional analyses are both time-consuming and expensive and thus inadequate to meet the increasing demand for rapid and cost-effective analyses in the cereal industry.

The aim of this thesis is to contribute to the development and demonstrate the potential of fast multivariate instrumental methods combined with chemometrics for the task of monitoring, handling and utilisation of quality variations in barley and wheat. This will be demonstrated by the seven included papers with examples of instrumental and chemometric applications to be used within plant breeding, handling, sorting and cereal process monitoring.

Most of the work presented in this thesis originates from the research projects: "Fast analysis methods for single seeds" (The Danish Cereal Network-B4, The Directorate for Food, Fisheries and Agri Business) and "Cascade Refining of European Wheat for Production of High-Quality Products for the Paper Industry"

(EU-FAIR CT96-1105). The aim of the first project was the development of combinatory analytical methods, mainly on a single seed basis. The aim of the latter project was a full-scale optimisation of the production chain from raw wheat grain to cationic modified wheat flour to be used in the paper industry, involving selection of wheat material, seed sorting, milling and chemical modification.

The thesis covers several aspects of different classical and instrumental analyses, all of which provide multivariate data which after interpretation with chemometrics are used for an improved quality characterisation in barley and wheat. Chapter 2 introduces the fast multivariate instrumental methods covering image analysis, hardness analysis and near infrared spectroscopy, all of which are adaptable to analysis of single kernels. Chapter 3 describes the chemometric tools which are utilised for full exploration of the large blocks of covariate multivariate data provided either by the fast instrumental methods or by classical analyses. In Chapter 4, the potential of the included papers is discussed in the context of the barley and wheat industry. Chapter 5 summarises the thesis with concluding remarks and perspectives on the presented results.

2. Fast instrumental methods

The cereal industry has been a pioneer in the application of fast spectroscopic analysis methods for monitoring of grain quality. Especially the development and use of near infrared spectroscopy has provided a tool for quality testing that is fast and cheap enough on a cost-per-sample basis for widespread use throughout the industry. Today, there exists a worldwide calibration network using near infrared transmittance (NIT) spectroscopy in combination with artificial neural networks (Büchmann *et al.*, 2001) for protein and moisture in various grains, underlining the success and potential of multivariate based spectroscopic techniques.

Such fast analytical methods are crucial in the primary cereal industry in order to secure raw materials of high quality for the following processing industries. The aim is to use a high and consistent quality grain input in order to optimise the processes and in order to avoid quality deviations in the end products. However, large variations in quality are often seen in the grain raw material. These grain quality variations are induced by genetic and environmental effects. The environmental differences can be seen between different geographical regions, differences between and within fields, and even between the seeds of a single plant. Thus, in order to secure cereal end products of high quality, fast analytical methods may assist at all levels of cereal production from breeding of new varieties, for trading/handling and cereal sorting, and for process monitoring in the food and feed processing industries. New fast instrumental methods in combination with chemometrics will thus assist in the comprehensive quality monitoring throughout the cereal industry.

In the following, a brief presentation will be given of the fast instrumental methods for cereal characterisation applied in this thesis: kernel morphology by image analysis, hardness analysis, and near infrared spectroscopy.

2.1. Kernel morphology by image analysis

Kernel size and shape parameters contain information relevant for the end-use quality. Automated image analysis has therefore become a promising analysis for the cereal industry. This technique has been used for discrimination between kernels of different species (Chtioui *et al.*, 1996), discrimination between wheat classes and varieties (Zayas *et al.*, 1986) and, used in combination with physical measurements, for variety identification (Zayas *et al.*, 1996). Berman *et al.* (1996) used the method for screening of flour milling yield in wheat breeding and Ruan *et al.* (1998) used image analysis in combination with neural networks for determination of fusarium scab in wheat.

In this thesis single seed image analysis was carried out by the use of a GrainCheck instrument (Foss Tecator, Sweden). Figure 2-1A shows the setup of this instrument. The system automatically acquires digital images (Figure 2-1B) of several hundred single seeds in a sample within a few minutes. These digital images are then used to estimate kernel characteristics such as kernel size, shape and colour of all the analysed kernels. The kernel characteristics are normally used in a discriminate analysis in order to detect impurities. In this thesis, the kernel characteristics are exported and used as a single seed "multivariate morphological fingerprint" (Figure 2-1C).

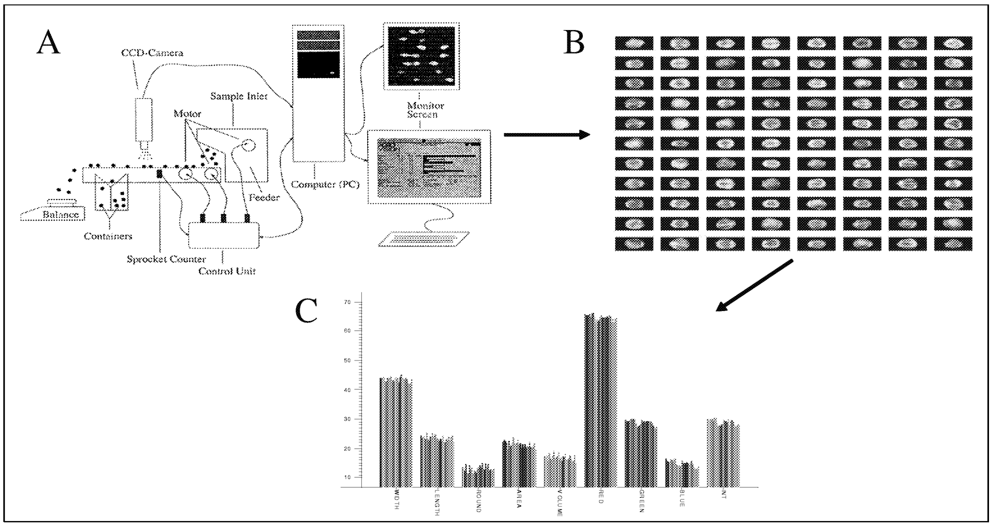


Figure 2-1. Image analysis of grains. A) Instrument setup of Foss Tecator GrainCheck. B) Images of wheat kernels acquired using the GrainCheck instrument, C) Histograms of nine morphological characteristics of single kernels estimated by the instrument

The single kernel readings are exploited in different ways in this thesis. In Paper 2 they are used for malting barley characterisation. The readings are represented as a kernel-to-kernel variability value of seed populations within each sample, either as histogram spectra of the different kernel characteristics or as sample mean and standard deviations. In Paper 6, each single wheat kernel is placed under the CCD camera and thereby true single seed recordings are used, retaining the identity of each kernel. Afterwards, other single seed analyses are added to the image data and used for an improved single kernel characterisation of wheat.

2.2. Hardness analysis

New instrumentation has also made automated single kernel hardness analysis possible. The Single Kernel Characterization System (SKCS) 4100 (Perten Instruments Inc., Reno, NV, USA) employed here is such an instrument for rapid, although destructive, measurement of single kernel hardness index (HI), weight (mg), diameter (mm) and moisture content (%) (Martin *et al.*, 1993). A rotating vacuum wheel picks up the individual kernels and deposits them one at a time into a weighing boat. After the weighing, the kernel passes down an inclined crescent where the diameter is measured and the kernel is then crushed between the crescent and a toothed rotor. A load cell measures and records the crush force-time profile for each kernel and its hardness index is calculated. The hardness index values are based on an algorithm that attempts to segregate wheats on a numeric scale on which hard wheats are forced toward an average value of 75, and soft wheats toward an average value of 25. The scale is similar to that used in the near infrared spectroscopic method (AACC 39-70A) for assessment of texture of bulk wheat samples. It has been shown that the SKCS provides fast information about conventional wheat quality factors such as NIR hardness, kernel size and test weight. It was, however, only found indicative for flour yield in milling (Osborne *et al.*, 1997; Ohm *et al.*, 1998). The SKCS measurements are normally conducted on 300 single kernels in a bulk sample in order to classify the sample into soft, hard or mixed wheat.

In this thesis the SKCS instrument is used for measurements of single seed hardness in both barley and wheat. In combination with the morphological data mentioned above, SKCS data are used as variability values (Paper 2) in seed populations (samples) for malting barley characterisation, and as true single kernel readings (Paper 6) for an improved characterisation of wheat.

2.3. Near infrared spectroscopy

Near infrared spectroscopy has proven to be a powerful tool in food and agricultural analysis. According to Williams and Norris (1987), "*Near infrared* *reflectance technology is the most practicable and exciting analytical technique to hit the agricultural and food industries since Johan Kjeldahl introduced the Kjeldahl test*". There are several reasons for the success of near infrared spectroscopy in agricultural analysis today:

- Near infrared spectroscopy is a most powerful tool for measurements of major chemical components in food and agricultural products such as moisture, protein, fat and carbohydrates.
- Near infrared analysis is fast, precise, non-invasive, easy to use, and only needs little or no sample preparation.
- The development of chemometric methods has allowed for an effective extraction of relevant information from the complex multivariate near infrared spectra.

In common with other spectroscopic methods, near infrared spectroscopy depends upon the principle that radiation interacts with matter to produce a response related to the physical and chemical properties of the sample, which through appropriate instrumentation can be displayed as a spectrum. The near infrared region covers the range of 780-2500 nm, between the visible and the mid infrared regions. The absorptions in this region correspond to overtones and combinations of the fundamental vibrational transitions in the mid infrared region involving, in particular, C-H, O-H or N-H due to the large anharmonicity of those vibrations involving the light hydrogen atoms. In near-infrared spectra the different constituents have broad overlapping peaks, so near infrared measurements have to be calibrated against samples with known chemical composition in order to extract the desired information. This is done with chemometrics.

In the beginning, near infrared spectroscopy was based on reflectance measurements performed on filter instruments, applying diffuse reflectance measurements on pre-ground grain samples for the measurement of moisture and protein (Williams and Norris, 1987). The prediction (regression) models were based on a few wavelengths (filters) calibrated to chemical analysis by multiple linear regressions (MLR). Later, monochromator-based scanning instruments were used, yielding continuous spectra, for example, from 1100-2500 nm. This gave a typical situation with a large number of variables (wavelengths) compared to the number of samples, which made multiple linear regression inadequate. Martens and Jensen (1983) therefore introduced the regression method Partial Least Squares

Regression (PLSR) for calibration of continuous near infrared spectra. The subsequent widespread use of continuous near infrared spectra thus became closely linked to the developments and successful use of multivariate regression methods such as Principal Component Regression (PCR) and PLSR (Martens & Næs, 1993; Fearn, 2001).

Near infrared radiation in the wavelength range of 1100-2500 nm will penetrate solid materials typically to a depth of only a fraction of a millimetre (Scotter, 1990). For this reason, heterogeneous material requires grinding before measurements. This, however, destroys the physical (structural) information which is contained in a near infrared spectrum of an intact sample. The use of the lower near infrared wavelength range (780-1100 nm) has allowed for transmittance measurements on whole grains due to more powerful radiation and lower absorbances in this range. By applying near infrared transmittance (NIT), a larger portion of the sample is irradiated and the homogenisation and packing of the sample is thus less critical.

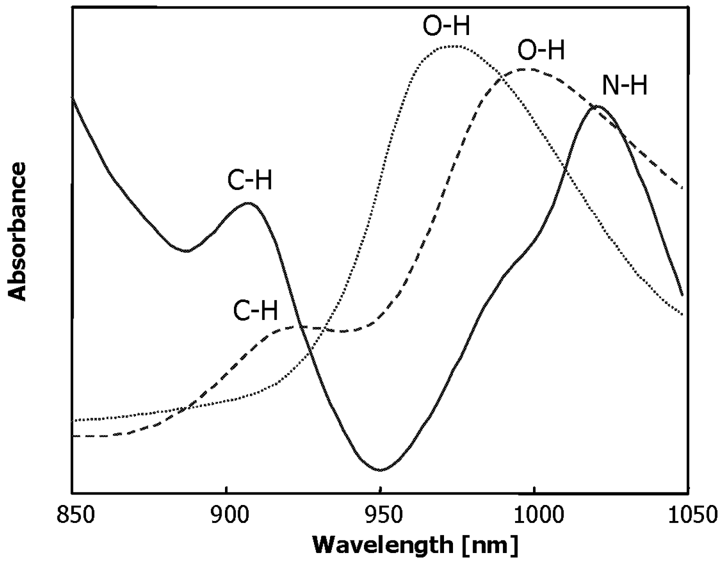


Figure 2-2. Near infrared spectra of the main components of wheat: gluten (—), starch (---) and water (···)

NIT spectra using an Infratec instrument from Foss Tecator cover the spectral region from 850 nm to 1050 nm containing primarily the second overtones of O-H (carbohydrates and water) and N-H (protein) stretching vibrations and the third overtone of the C-H (fats) stretching vibration. Near infrared spectra of gluten, starch and water in the spectral region of 850-1050 nm are shown in Figure 2-2 with assignments of the second and third overtones of fundamental N-H, C-H and O-H stretching corresponding to the expected band in this region.

This underlines the repetitive holographic nature of near infrared spectra, as the relevant information found in this narrow spectroscopic window is a consequence of the fundamental vibrations in the mid infrared region.

Near infrared spectroscopy has become a widely used method in analysis of cereals and cereal-based products, as extensively reviewed by Osborne *et al.*, (1993), Williams and Norris, (1987), Williams, 2002 and Meurens and Yan (2002). Today, NIT instrumentation for whole kernel analysis on bulk samples is used worldwide in combination with advanced multivariate regression methods (Büchmann *et al.*, 2001), and it has nearly the status as "the new reference analysis" for protein and moisture in grains. New instrumentation in near-infrared spectroscopy has made single seed analysis possible, and as briefly reviewed in Paper 6, the method have been reported on different types of grains. Most of the single seed applications on cereals has been dedicated to wheat for classification purposes (Dowell, 2000; Delwiche and Massie, 1996; Wang *et al.* 2002) and for the prediction of protein content (Delwiche, 1995; Delwiche, 1998; Delwiche and Hruschka, 2000) and hardness (Delwiche, 1993).

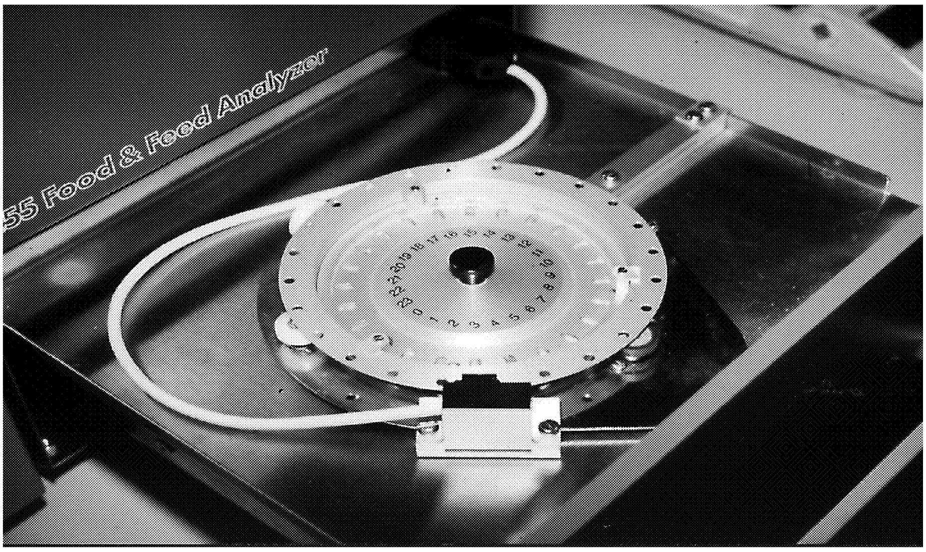


Figure 2-3. Single seed adapter in a NIT Infratec 1255 Food and Feed Analyzer, making it possible to measure near infrared transmittance (NIT) spectra of single seeds

Near infrared spectroscopy has been exploited in different ways in this thesis for a range of purposes. Near infrared reflectance (NIR) spectroscopy and NIT spectroscopy have been applied on whole barley and malt kernels in bulk (Paper 3, 4 and 5). NIR spectroscopy was used to analyse barley flour (Paper 5), malt flour (Paper 3) and wheat flour (Paper 7). In Paper 6, the potential of single seed NIT spectroscopy for wheat characterisation is demonstrated. For these measurements a single seed adapter (Figure 2-3) attached to a NIT Infratec 1255 Food and Feed Analyzer (Foss Tecator) was employed. Each single kernel was manually placed in the pockets and NIT spectra in the range of 850-1050 nm were measured and used for physical and chemical characterisation of single wheat kernels (Paper 6).

3. Chemometric evaluation tools

The fast instrumental methods used in the work presented in this thesis produce large blocks of multivariate data. Chemometric analysis of such data allows for a pattern recognition strategy in which numbers are transformed into graphic display supporting the interpretation. Complex multivariate data sets are thereby explored with a minimum of pre-assumptions, by mathematically reducing their dimensionality into fundamental underlying factors, the nature of which can subsequently be explained by validation against prior knowledge.

The terms chemometrics and multivariate data analysis are often used synonymously, where the latter refers to multivariate data analysis applied in any field of science, while chemometrics is the application of multivariate mathematics to efficiently extract maximum useful information from chemical data. Chemometrics was inspired by social sciences such as psychometrics and econometrics dealing with real world multivariate data. These sciences have been aware of the importance of presenting data in a user-friendly visual form for interpretation.

In this thesis Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) are used. PCA is used for decomposition of a single multivariate matrix for interpretation and data overview. PLSR is used for two-block multivariate regression models. The following chapter includes a description of the principles of PCA and PLSR along with a brief discussion on validation, outliers, preprocessing of data, and variable selection.

3.1. Principal Component Analysis

Principal Component Analysis (PCA) (Wold *et al*., 1987) is a powerful exploratory technique for compression of large multivariate data sets for classification purposes. The earliest approaches towards PCA were taken by Pearson (1901) and Hotelling (1933). As the name indicates, the PCA algorithm finds the main directions in a multidimensional data set by creating orthogonal principal components whose linear combinations approximate the original data in a least-square sense. The original data matrix (**X**) is decomposed into a score matrix (**T**) and a loading matrix (**P**) and the residuals are collected in a matrix (**E**):

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

Only a limited number of components are relevant for describing the information in **X**. The scores (the **t**'s) contain information about the samples and the loadings (**p**'s) contain information about the variables. The loadings are common to all samples, and the scores specify the amount (concentration) of the common loadings within each of the samples. Patterns and clusters of the objects are easily represented in the form of scatter plots of the scores (score plots) by exploring different combinations of principal components as axes. Figure 3-1 shows an example of a PCA score plot of scatter-corrected single seed NIT spectra of wheat (w) and barley (b) kernels.

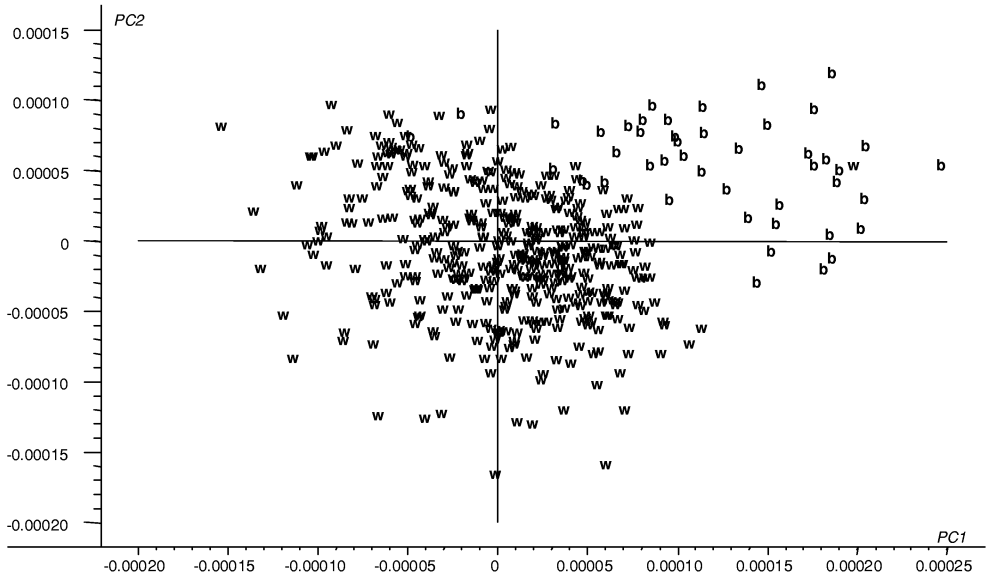


Figure 3-1. PCA score plot (PC1 versus PC2) of scatter-corrected NIT spectra of single wheat (w) and barley (b) kernels

The score plot reveals an almost clear differentiation between the wheat and barley kernel spectra, however with a "wheat outlier" among the barley spectra and a few "barley outliers" among the wheat spectra. This is thus an unsupervised "let the data speak for themselves" approach in which expected and unexpected trends in complex data can be found.

In this thesis, PCA results are presented in Papers 1, 2, 4 and 5, but in the remaining investigations PCA has also been used as an initial analysis for detection of data trends and outliers prior to the more straightforward regression models using Partial Least Squares Regression.

3.2. Partial Least Squares Regression

Partial Least Squares Regression (PLSR) (Wold *et al.*, 1983; Martens and Jensen, 1983; Martens and Næs, 1989) is a predictive two-block regression method based on latent variables and is applied to the simultaneous analysis of two data matrices. The purpose of the PLSR is to build a linear model between a desired **y** characteristic, e.g. protein content, from easily obtainable **X** data, e.g. near infrared spectra. While PCA is unsupervised, PLSR is supervised in the sense that it focuses on finding information in **X** relevant for describing one or several *a priori* defined **y** characteristics.

In matrix notation we build the linear model:

$$\mathbf{y} = \mathbf{Xb} + \mathrm{e}$$

where **b** contains the regression coefficients that are determined during the calibration step and "e" the residuals (model errors, noise etc.). In a PLSR calibration, a multiple linear regression model is built between the significant scores and the **y**. Compared to the PCA scores, the significant PLSR scores are found in a slightly different way, taking into account the variation in **y** during the decomposition of **X**, i.e. the covariance between the scores in X and **y** is maximised.

The regression coefficients **b** computed during the calibration, together with a new **x** (e.g. a new measured NIR spectrum), is then used for prediction of the desired characteristics (y_{pred}) of a new/future sample:

$$\mathbf{x'}_{\mathrm{new}}\mathbf{b} = y_{\mathrm{pred}}$$

One important feature of PLSR is the ability to model covariate data which is in contrast to Multiple Linear Regression (MLR). MLR is the classical way of building a regression model using several **X**-variables, but MLR is designed for independent variables, and does thus not cope well with covariate data.

Figure 3.2 shows an example of a NIT based PLSR model for prediction of protein in single kernels of wheat (w) and barley (b) (the same spectra as used in the PCA score plot in Figure 3-1). This example shows that when applying an exploratory PCA strategy, it is possible to differentiate between wheat and barley spectra (Figure 3-1), and at the same time, when applying a supervised regression method like PLSR, it is possible to build a calibration model for a common classical analysis – protein – across both wheat and barley kernel spectra (Figure 3-2).

In this thesis PLSR has been applied in Papers 1 to 6.

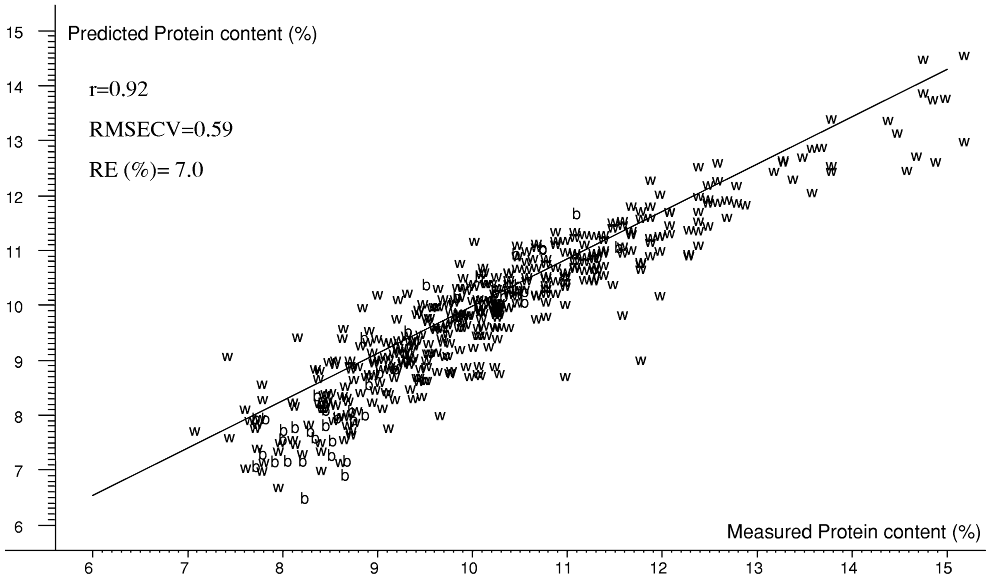


Figure 3-2. Predicted versus measured protein content of a 6 component PLSR model based on scatter-corrected NIT spectra of single wheat (w) and barley (b) kernels. The correlation coefficient (r), cross-validated prediction error (RMSECV) and relative prediction error (RE) are given. These terms are described below.

3.3. Validation

3.3.1. Model validation

Model validation is of great importance in chemometrics in order to provide information on possible outliers, number of latent factors to include in the model, and prediction errors. There are basically two ways to validate, namely test-set validation, and in the case of small data sets, cross-validation. Test-set validation requires two independent data sets representative of the sample population, where the model is built on one data set and tested on the other. Cross-validation is used when too few samples are available to obtain an independent test set. In cross-validation the data set is either randomly or orderly divided into a number of segments of one or more samples. Each segment is successively excluded and used for testing the model based on the remaining samples. In this way all samples are used for estimation of the model, and all samples are excluded from the model in order to validate this effect on the remaining model.

The calibration models in this thesis were mainly validated by cross-validation (Papers 1-6), but test-set validation was also applied (Paper 6).

The performance of a regression model is normally evaluated by its prediction error in terms of root mean square error of prediction (RMSEP) for true test set predictions, and root mean square error of cross-validation (RMSECV) for cross-validated models, as given below:

$$\text{RMSEP or RMSECV} = \sqrt{\frac{\sum_{N}\left(y_{\text{pred}} - y_{\text{ref}}\right)^2}{N}}$$

where y_{pred} is the predicted value using either test-set or cross-validation, y_{ref} is the reference value, and N is the number of samples. In addition, the correlation coefficient (r) between the reference value (y_{ref}) and the predicted value (y_{pred}) and the number of latent factors (model complexity) are used for model evaluation (see example in Figure 3-2).

Outlying samples can deteriorate a multivariate model and should normally be eliminated before modelling. Samples may, however, be classified as outliers for several reasons. Outliers where the data are not valid should, of course, not be included in a calibration model, while outliers with, for instance, extremely high or low concentration of the constituent in question may improve the model when included by extending the limits of the future predictions. Automatic outlier detection is applied in the NIT Infratec instruments (Lindholm, 2002). When a new sample is analysed and predicted by the multivariate model in the instrument, outlying behaviour will be displayed automatically and the problem categorised (residual, leverage, sub-sample deviations, out of range or temperature deviations). Hereby, the operator can solve the problem or reject the analysis result and save the sample for later inclusion in the model. From an analytical point-of-view, outliers are often problematic, but from a breeding point-of-view, outliers may be a potential mutant or transformant, as exemplified in Paper 5 where a sample covering a gene for changed protein composition is a clear outlier in a PCA on a normal barley population.

3.3.2. "Practical" validation

When evaluating the usefulness of a chemometric model it is not always sufficient to consider the model performance parameters (RMSECV, RMSEP etc.). It is often necessary to evaluate the model with new independent measurements both in the context in which it was developed and in the context in which it is to be used in the future. An easy way to evaluate the usefulness of a prediction model is to compare the prediction error to the variation in the parameter in question. Williams and Sorbering (1993) suggested the term Relative Prediction Deviation which is defined as the standard deviation in the population of prediction samples divided by the standard error of prediction (SEP).

In this thesis (Papers 1, 2, 3, 4 and 6) a similar approach is used, simply by calculating the relative error (RE) in percent as:

$$\mathrm{RE} = \left(\frac{\mathrm{RMSEP\ or\ RMSECV}}{y_{\max} - y_{\min}} \right) * 100$$

where the $y_{\max}$ is the highest reference value and the $y_{\min}$ is the lowest reference value of the **y** parameter in question. The RE value is unit-less and can be used when calibrations for different parameters are to be compared. Low RE indicates good calibration models which might be used for a precise quality control and in cereal trade, while calibration models with medium RE might be used for a rough selection in plant breeding. Models with high RE's are not suitable for practical use.

The performance of a chemometric model is highly dependent of the accuracy of the input data. Thus when validating a model, it is important to consider the uncertainty of the measurements used in the model, for example, considering NIR spectra and the protein content determined by Kjeldahl in a PLSR calibration. A comparison between the uncertainty of the input data and the prediction error will thus indicate if there is room for improvement of the model. Normally, the quality of the data can be improved by replicate measurements, for example, by several spectral recordings on the same sample or several sub-sample analyses of protein content on each sample. However, this is not possible when applying destructive single kernel analysis, such as Kjeldahl protein, since only one analysis can be performed on a kernel. In Paper 6, an alternative way was suggested in order to circumvent this problem by mathematically simulating replicate measurements by averaging data across single kernels that are nearly identical in reference value. By comparing this approach on a model for protein content (accurate reference method) and a model for hardness (less accurate reference method) on the same spectra (**X**), it was possible to indicate that the inferior predictive ability of the hardness model was due to inaccuracy in reference method of single seed hardness determination (SKCS 4100).

From a practical point of view it is also important to consider the potential advantages, such as speed of analysis, when selecting the predictive model to be applied in the laboratory. For instance, when predicting malting barley quality (Papers 2 and 3 and Chapter 4), the fast methods can be applied either on the raw barley, on the micro-malt or on the final wort. For early selection in malting barley breeding programs one would probably allow a higher RE, when predicting malt quality on the level of the barley raw material compared to measurements on the malt, since the time of analysis (including micro-malting) is considerably lower for the raw barley measurements. Thus, the choice of application is a compromise between required accuracy of the predictive model, on one hand, and measuring ease and sample throughput on the other hand.

A final evaluation of a given grain lot is normally based on several parameters simultaneously as well as on experience and accumulated knowledge of the malster/brewer or miller/baker, in which each quality parameter is evaluated according to a target or target range. The significance of the single analyses are then summarised in a total evaluation, to be used for final acceptance/rejection and price determination of a given grain lot. When applying validation in this context, the ranking of the expert evaluation is recommended as the "reference method" on the basis of which validation should be made, as shown in the simplifying approach for malt quality evaluation in Paper 4.

The ultimate validation of any multivariate application is, however, not done until the models have been implemented in the industry, and have rendered reliable and useful results. The application of NIT for determination of protein and moisture in a global network is an excellent example of such an "industrial validation" (Büchmann *et al.*, 2001).

3.4. Chemometric pre-processing of near infrared spectra

In addition to the useful chemical and physical information, multivariate spectral data contain information which may be irrelevant to the parameter in question. This may be due to instrumental drift or scatter because of different physical and optical properties of the measured samples. In both cases it will, in a systematic way, influence the level of signal. In this way, linear modelling becomes difficult, since the relevant variation is influenced by irrelevant, but systematic, variation.

In order to eliminate or at least reduce the non-relevant spectral information, pre-processing of the spectra can be performed, which will often lead to simpler and more robust regression models. Several pre-processing methods have been developed, e.g. Multiplicative Signal Correction (MSC) (Martens *et al.*, 1983; Geladi *et al.*, 1985), Piecewise Multiplicative Signal Correction (PMSC) (Isaksson and Kowalski, 1993), Standard Normal Variate (SNV) (Barnes *et al.*, 1989) and derivatives (Savitsky and Golay, 1964). Empirical evidence has accumulated for especially spectral derivatisation and MSC as successful techniques. In this thesis derivatisation was applied on NIT spectra in Papers 4 and 5, MSC was applied on the NIR spectra in Paper 5, and the SNV technique was applied on NIR reflectance spectra in Paper 7.

The single seed protein system studied in Paper 6 uses a combination of the second derivative followed by MSC, which was originally suggested by de Noord (1994) and used by Delwiche (1995) for the same purpose. This double transformation naturally calls for the development of more general and powerful pre-transformations, as described in collaboration with Martens *et al.* (2002) and Pedersen *et al.* (2002) (Appendices 2 and 3) proposing slight different versions of the Extended Inverted Signal Correction (EISC). These two new data transformations are based on the inverted version of MSC; Inverted Scatter Correction (ISC) (Helland *et al.*, 1995) with chemical and physical extensions. The MSC and ISC have in common that they are relatively simple, and that they can be applied without *a priori* knowledge about the samples. The core of MSC and ISC involves correction of each spectrum in a set of related samples towards an "ideal" spectrum where the physical scattering has been removed. Linear regression between the input spectrum and the ideal spectrum is employed to estimate the correction coefficients *a* (additive effect) and *b* (multiplicative effect). Theoretically, *a priori* knowledge about the samples and their spectra can enhance the performance of the MSC methods. Martens and Stark (1991) included information about the major analyte spectra in the MSC estimation of *a* and *b* in their "Extended Multiplicative Signal Correction" (EMSC). In Martens *et al.* (2002), different *a priori* MSC/ISC extensions were applied to NIT spectra (850-1050 nm) of five binary mixtures of gluten and starch. Different packing of the cuvette as well as different samplings induced considerable scatter in the NIT spectra, as shown in Figure 3-3A.

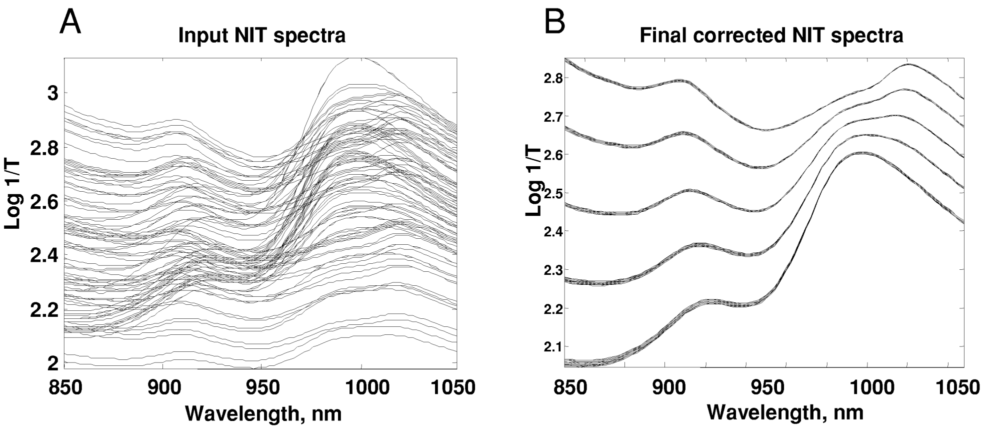


Figure 3-3. NIT spectra (850-1050 nm) of 5 powder mixtures of gluten and starch shown as the raw spectra (A) and after full EISC correction (B) (modified from Martens *et al.* 2002).

These spectra were drastically improved by extending the ISC pre-processing model with *a priori* chemical (analyte and interferences) and physical (wavelength dependence of scatter) spectral information. Both the EMSC and EISC were able to isolate and remove additive, multiplicative and wavelength-dependent effects of light scattering so effectively that the pre-processed NIT spectra of the powder mixtures appeared as if they represented NIT spectra of five "transparent liquid solutions" (Figure 3-3B).

However, in spectroscopic analysis of complex samples it is often difficult to include *a priori* analyte information. This is also the case in the single seed wheat study (Paper 6) in which the NIT spectra are complex combinations of different scatter effects together with information of chemical constituents included in the kernel matrix. Figure 3-4A shows the raw NIT spectra of the 415 calibration wheat kernels.

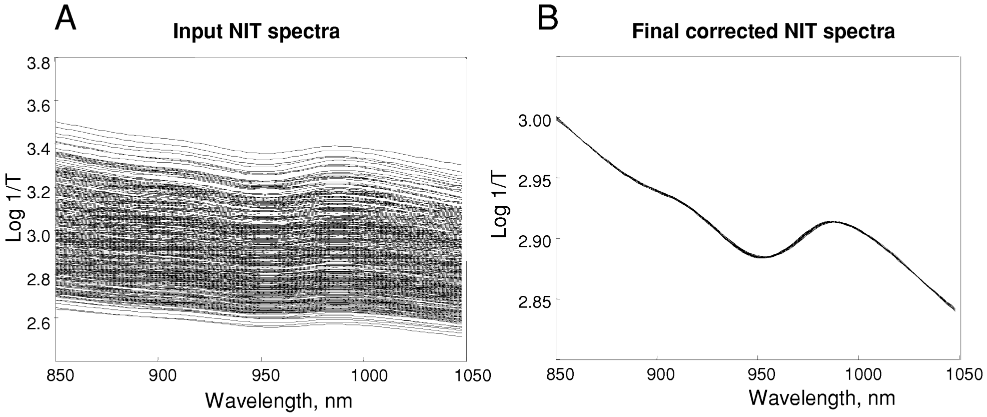


Figure 3-4. NIT spectra (850-1050 nm) of 415 single wheat kernels shown as the raw spectra (A) and after full EISC correction (B) (modified from Pedersen *et al.* 2002)

Although the application of EISC was limited only to include physical extensions (Pedersen *et al.*, 2002), it proved to be a powerful pre-transformation. The EISC-corrected NIT spectra of the 415 single wheat kernels (Figure 3-4B) performed equally well as the two-step second-derivative followed by MSC used by Delwiche (1995) and in Paper 6 for the prediction of protein in single seed of wheat.

3.5. Variable selection

Chemometrics favours the use of principal components or latent factors extracted from all the recorded multivariate variables. For several reasons it can, however, often be useful to focus on specific variables or variable regions/intervals in the multivariate modelling. First of all, the modelling (e.g. PLSR modelling) often improves when focusing on relevant spectral regions and leaving out interfering and noisy regions. Such variables or variable regions could also be used for development of low cost, high speed instruments based on, for instance, filters. For interpretation purposes variable selection is also valuable in order to validate “hot spots” in the modelling with spectroscopic knowledge of the absorbers in the system.

There are several methods for variable selection such as Principal Variables (Höskuldsson, 1994), forward stepwise selection and genetic algorithms (Leardi, 2001). The Interval PLS (iPLS) was proposed in collaboration with Nørgaard *et al.* (2000) (Appendix 1). This variable selection method provides a new graphically oriented local modelling procedure for use on spectral data and is an interactive extension to PLS, which develops local PLS models on equidistant subintervals of the full-spectrum region. Facilitated by graphic display, these subinterval models are compared with regard to prediction performance. The best equidistant interval is subsequently optimised by small adjustments of the interval limits by shifting the interval position and by changing the interval width (both symmetrical and asymmetrical). On a data set representing NIR spectral data on 60 beer samples correlated to original extract, the iPLS method proved to represent a sound compromise combining data reduction with spectral localisation, while still being able to utilise the multivariate advantage (Nørgaard *et al*., 2000).

In Paper 3 the iPLS algorithm has been applied on NIR and NIT spectra of malt samples in order to improve the prediction of extract, nitrogen in malt, modification and ß-glucan in malt and wort. For all the NIR (400-2500 nm) models, considerable model improvements in terms of lower prediction errors and lower model complexity were seen when limiting the analysis to informative spectral regions found by the iPLS algorithm. In contrast, no improved spectral region could be found in the NIT spectra. This could be due to the narrow spectral range (850-1050 nm) together with the fact that this region mainly represents broad and overlapping peaks of second and third overtones and combinations hereof.

An extension of the iPLS algorithm called synergy iPLS (si-PLS) has been developed in-house by Associate Professor Lars Nørgaard (www.models.kvl.dk) (method unpublished). This algorithm combines all possible combinations of intervals in order to find the combination of informative spectral regions which gives the lowest prediction error. The si-PLS was applied in Paper 5 in order to interpret the link between NIR information and endosperm constituents in a material representing wild types and a proteome-altering gene mutant in barley. The results were fully interpretable with regards to gene classification. These results, however, together with the results in Paper 3, demonstrate that near infrared spectra contain repetitive overlapping chemical information throughout the spectrum, so the relationship between selected regions and chemical/physical compositions are not always fully interpretable.

Martens and Martens (2000) recently proposed the modified Jack-knife validation for uncertainty estimation of **X** variables in a PLSR. In this method the regression coefficients of all the sub models in a cross-validated PLSR are used to estimate the uncertainty of each variable regressor. It is thus possible to inspect each of the variables in **X** used in the PLSR with respect to both importance (magnitude) and uncertainty. The method was employed by Westad and Martens (2000) on NIR spectra and in this thesis for the purpose of interpretation of malting barley quality parameters (Paper 1) and interpretation of the link between barley kernel morphology, hardness and malting quality (Paper 2).

4. Fast instrumental and chemometric applications in the cereal industry

The modern processing technologies in the milling, baking, malting and brewing industries require grain raw material of high and consistent quality. The aim is to optimise the raw material and the processes in order to avoid unwanted quality deviations in the end products. Large variations in quality are, however, seen in the raw grain material, focusing here on barley and wheat. These grain quality variations are induced by both genetic and environmental differences and their interactions. The environmental quality differences are due to differences in weather conditions, soils characteristics and growing practices. These environmental differences are seen on different levels with regard to geographical areas, fields, part of fields and even reflecting single seed differences within a single plant.

This chapter will demonstrate that fast multivariate instrumental methods in combination with chemometrics can be used in plant breeding (section 4.1), in handling and exploiting environmental quality differences (section 4.2) and in quality control in cereal processing (section 4.3).

4.1. Fast screening methods in plant breeding

Plant breeding involves selection and crossing of parents that carry required characteristics followed by evaluation of the progeny to identify which lines have inherited the desired characteristics, combining seed quality characteristics with grain yield and disease resistance. This activity involves screening of very large populations of early generations in order to eliminate undesired lines, so that the most promising material can be carried to more advanced stages. Thus, fast screening methods in breeding need not be completely accurate, but should have the capability to classify the material into good, acceptable and rejected categories.

In the following, examples will be given of how fast multivariate methods in combination with chemometrics can be utilised in plant breeding.

4.1.1. Exploring genotypic and phenotypic variation using NIR spectroscopy for screening of nutritional protein value in barley

Rapid screening methods for chemical composition and identification of specific genes and gene effects are important tools in plant breeding and biotechnology. Paper 5 introduces a new idea in which pleiotropic effects of a gene mutant in an isogenic background such as in barley and wheat can be detected by spectroscopy and chemometrics. Recently, NIR spectroscopy has been successfully applied to detect the phenotypic effects of wheat-rye chromosomal translocation (Delwiche *et al.*, 1999) and for the identification of waxy wheats (Delwiche and Graybosch, 2002). Wang *et al.* (1999) used NIR for prediction of the number of dominant R alleles (coding for red pigmentation in the seed coat) in single wheat kernels, and Campbell *et al.* (2000) investigated the use of NIT for classification of starch mutants in corn.

The investigation described in Paper 5 involves 125 samples of normal barley lines and the regulator gene *lys3a* in different genetic backgrounds. The *lys3a* is a high-lysine mutant originally found at the Risø Laboratory, Denmark by a dye-binding method involving acilane orange developed by Munck (1992). This gene drastically changes the proteome as displayed by 2-D electrophoresis, resulting in drastical changes in amino acid profile. Through pleiotropic effects it also changes the whole cell machinery which is important for the synthesis of chemical constituents such as starch, fibre and fat. By applying exploratory PCA on NIT spectra of the samples, a clear clustering was seen between the normal barley lines and the *lys3a* recombinants. This shows that non-destructive NIT analysis on whole grains could be used in plant breeding to select mutants such as *lys3a*, as the dye-binding method was originally used. Furthermore, a PCA on NIR spectra shows a clear clustering both with regard to genetic (normal versus *lys3a*) and growing conditions (field versus green house). These spectral differences were validated to amino acid and chemical analysis using PLS (full-spectrum PLS, iPLS and si-PLS (see Chapter 3)). The PLSR models for protein-N and amide-N did not only rely directly on the protein information in the spectra, but also on information concerning other components such as fibre and fat differences, exploiting the pleiotropic effects of the gene. It was also demonstrated that the gene effect could be represented as a “spectroscopic signature” characteristic for the genotype. The

NIR method thus provides much more holistic information compared to the previously used dye-binding method.

In barley lines with the high-lysine gene (*lys3a*) the relationship between the concentration of essential amino acids in the protein and the total protein content is changed. In normal barley there is a negative relationship, i.e. the higher the protein content, the lower the concentration of the 8 essential amino acids (e.g. lysine) in the protein. In the high-lysine mutants this relationship is straightened out, so that the concentration of lysine is independent of the total protein content (Paper 5; Munck, 1992). It is hereby possible to increase the total amount of essential amino acids and decrease the amount of non-essential amino acids without increasing the total protein content.

The feeding industry has not traditionally been interested in variation of nutritional quality in cereals. Recently, more attention has been paid to the nutritional value and digestibility of cereals for pig and poultry feeding in order to improve the efficiency of the animal production and in order to reduce the leaching problems of nitrogen and phosphorus from the intensive animal productions. In Paper 5 a laboratory method for alkali volatile nitrogen (amide-N) is described, and a "amide-N:total N ratio" is proposed, which is shown to correlate (negatively) to the concentration of lysine (and other essential amino acids) in the protein. The ratio clearly separates the normal and *lys3a* barley lines. On a limited material it was possible to predict this ratio using NIR (r=0.96, RMSECV=0.76) and there was indication of a possibility to predict the lysine concentration in the protein (r=0.95, RMSECV=0.24) (Paper 5). NIR seems, therefore, to constitute a potentially useful screening method in breeding of barley with improved nutritional efficiency of protein.

4.1.2. Fast multivariate characterisation of malt quality

The quality of barley and malt used in the modern malting and brewery industry is important from both a process technological and an end product quality point-of-view. A full characterisation of malting barley quality includes a range of reference parameters reflecting physical, chemical and functional properties. These analyses are produced by expensive, time-consuming and destructive methods and the need for screening methods in barley plant breeding is therefore obvious in order to facilitate a higher sample throughput. The perfect instrument would be a non-destructive screening measurement, which by analysing the ungerminated barley samples would be able to predict all barley, malt and wort parameters. The conversion of barley into malt involves complex interactions between biochemical and structural parameters in the germination. The basic physical and chemical composition is assessable directly from the barley raw material, while the physiological and biochemical changes during the germination are difficult to predict from the intact barley kernels.

On a data set of 50 barley samples grown at two locations in Denmark, the use of barley kernel morphology and hardness for malting barley characterisation was been investigated using the GrainCheck and the SKCS 4100 (Paper 2). Both instruments are fully automated and provide single kernel data of 250 kernels within a few minutes, making them of interest for screening purposes in malting barley breeding programmes. High barley grain homogeneity is of great importance to secure uniform germination and thereby secure an evenly modified malt. The GrainCheck and SKCS 4100 automatically provide seed homogeneity data for each sample with regard to morphological and hardness characteristics. These data have been exploited in two different ways, either as means and standard deviations or as appended histogram spectra of 250 seeds from each bulk sample and utilised for the prediction of 13 malting barley parameters. Thus it was investigated whether the single seed information from batches provided additional information compared to sample average for barley quality characterisation.

It was shown possible to obtain reasonable PLSR models, based on barley kernel morphology and hardness, for the structural and physical part of the malting quality complex associated to malting modification. Not surprisingly, however, it was impossible on the level of the barley seed to model the biochemical parameters associated to germination and enzymatic power. The utilisation of the single kernel advantage from the two applied instruments only seemed to provide additional information regarding malt homogeneity where the hardness homogeneity within the samples was the most important variable. The prediction error of this model, however, was too high for practical use (Paper 2). The SKCS Relative Hardness Index (RHI) is by far the most important of the investigated variables for describing the malting performance. The additional use of the morphological data, as acquired by fast non-destructive image analysis, also reflects some malting quality information by improving the calibration models based on RHI alone. The brightness of the kernels is by far the most important morphological variable here.

To be able to compare the potential of these PLSR models based on morphological and hardness data with models based on near infrared spectroscopy, the exact same sample set was analysed using near infrared spectroscopy on whole kernels (both barley and malt). Near Infrared spectroscopy is widely used for the prediction of various barley and malting quality traits in the cereal industry.

NIR spectroscopy on bulk samples has been applied on whole barley kernels (Williams and Sobering, 1993; Halsey, 1987; Roumeliotis *et al.*, 2000), ground barley (Allison, 1989; Sinnaeveg *et al.*, 1994; Henry, 1985; Henry, 1985a; Faccioli *et al.*, 2000; Szczodrak *et al.*, 1992) and ground malt (Sinnaeveg *et al.*, 1994; Henry, 1985) for a range of malting quality traits. NIT spectroscopy has also been successfully applied on whole barley and whole malt (Williams and Sobering, 1993, Sinnaeveg *et al.*, 1994) for malting quality predictions, and national and international calibration networks for the prediction of barley and malt quality are currently is based on this technique (Day, 1999; Büchmann *et al.* 2001).

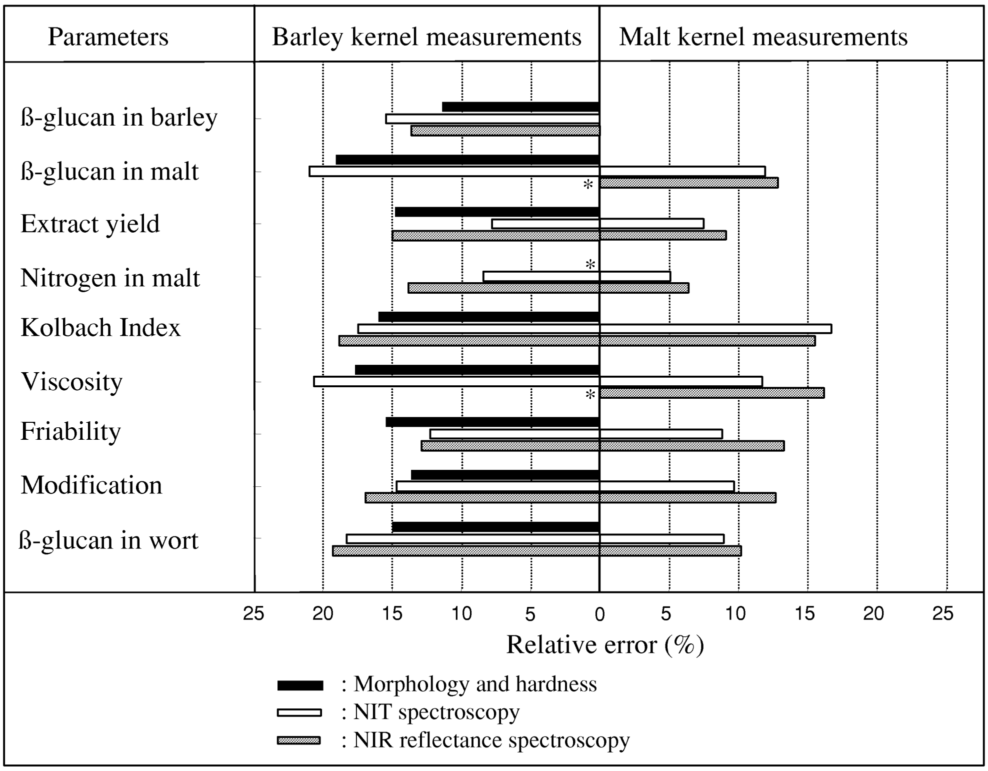


Figure 4-1. PLSR relative prediction errors using morphology and hardness, NIT or NIR reflectance spectroscopy for nine malting barley quality parameters. *: model excluded due to too poor predictive ability

In this thesis, the NIT spectra are recorded using a NIT Infratec 1225 Food and Feed Analyzer (FossTecator, Höganäs, Sweden) and the best models, shown as white bars in Figure 4-1, are either based on raw (Paper 3) or second derivative corrected spectra (unpublished, model details not shown). The NIR spectra were recorded using a NIRSystems 6500 (Foss NIRSystems, Silver Spring, Maryland, USA) with a coarse sample cell and the best models, shown as grey bars in Figure 4-1, are either based on raw (Paper 3) or MSC-corrected spectra (unpublished, model details not shown). The results of the PLSR relative prediction errors (RE) using mean and standard deviations of morphology/hardness data on barley kernels (Paper 2) are also given in Figure 4-1. The predictions of diastatic power, wort colour, nitrogen in wort and homogeneity were rather poor using any of the three methods, and are therefore excluded from the figure. It is seen that models based on morphology and hardness data perform well for the prediction of the nine quality parameters shown. For most of the parameters, the prediction results are even slightly better than the results of the prediction models based on NIT and NIR spectroscopy on barley.

A comparison of the near infrared measurements on barley (left) and malt (right) kernels shows, as expected, that the models based on malt are superior to models based on barley. This underlines the importance of the germination properties in the barley during malting which cannot be assessed by spectroscopy on the barley itself. Moreover, it is seen that the NIT measurements for most of the parameters are superior to the full-spectrum NIR measurements. However, Paper 3 showed that when focusing on the relevant part of the NIR spectra, these PLSR models can be improved considerably; underlining that both NIR and NIT on whole kernels can be used.

4.1.3. Multivariate quality assessment systems

As reviewed by Siebert (2001), chemometric methods are being more and more used in the brewing area at all levels from barley to final beer. Within the area of barley and malt, PCA has been demonstrated to be a powerful tool to overview and simplify several barley and malt parameters into underlying factors (Jacobsen, 1980; Munck, 1991; Lonkhuijsen *et al.*, 1998). PCA and PLSR were used in Paper 1 for an exploratory study of micro-malting data of spring and winter barley samples. The underlying principal components form more or less interpretable functional factors, patterns of which are evaluated by loadings plot (Paper 1). The corresponding scores (representing the samples), visually presented as score plots, revealed clusters and trends and was used in evaluating the tested genotypes and their interaction with growing location (Paper 1).

The use of PCA for a more supervised ranking of malting barley quality has been proposed (Nielsen and Munck, 2000). The idea was to construct an "ideal sample" where an "optimal" value of each of the included parameters has been assessed. The principle is to include this ideal quality profile as a new artificial sample in a data set of samples to be evaluated. A PCA is then applied and the scores are used to calculate the distance between the ideal and the other samples. The shorter the distance, the better the quality of the sample. This approach was to some extent successful; however, the ranking was not in full agreement with an expert ranking.

Several indexes have been proposed for calculation of overall malt quality. Most of these indexes are based on weighted linear combinations of a limited number of parameters (Molina-Cano *et al.*, 1986; Molina-Cano, 1987; Madre, 2002). A corresponding approach was tested (Nielsen, 2001) in which the distance from an "ideal sample" based on a weighted linear combination of ten malt quality parameters was calculated. This approach turned out to give a reasonable index which furthermore was predictable by NIT spectra of the malt sample (r=0.88, RE=10%).

Both the PCA approach and weighted linear combination approach assumes a linear relationship between the level of a certain parameter throughout the parameter range and the degree of acceptance by the end-user. This may not be appropriate. Furthermore, the above approaches do not take into account that the level of a certain parameter can be completely out of specification and thereby unacceptable. The idea presented in Paper 4 will allow a simple way of treating the above problems. This approach in based on expert knowledge, in which fuzzy membership functions for each of the parameters are defined. A fuzzy membership function (exemplified in Paper 4) defines to which degree the quality parameter is acceptable on a scale from zero (unacceptable) to one (optimal). The memberships (one for each quality parameter) are combined into an overall quality index (OQI) by means of a simple weighted addition. The weights are defined on the basis of expert evaluation of the importance of each parameter on a scale from 1 (less important) to 10 (very important).

This fuzzy logic approach was employed on malt quality data of the same 50 samples as discussed in Papers 1, 2 and 3. The calculated OQI's of all the samples were thoroughly validated and found to be a sound index reflecting prior knowledge and expectations of the tested malt samples. Eleven malt quality parameters are hereby combined into one number to be used for overall ranking of malt samples. In Paper 4 it is furthermore shown that a reasonable PLSR calibration can be built between the OQI and bulk NIT spectra of the malt samples (r=0.88, RMSECV=11, RE=15%). This demonstrates that the physical and chemical information in the NIT spectra contains most of the information in the OQI. The fact that the NIT spectra do not perfectly predict the OQI's may be due to the fact that some of the parameters (wort colour, diastatic power, soluble N in malt and homogeneity) used in the OQI are not predictable by NIT spectra of the malt samples by a traditional single parameter PLSR calibration, as shown in Section 4.1.2. Nevertheless, it is demonstrated that fast multivariate near infrared spectra may be used in a more holistic way for characterising multivariate malting barley quality to be used in plant breeding.

4.2. Handling and utilisation of grain quality variations

As mentioned earlier, quality variation can be seen between geographical regions, between different fields, and on the macro and micro level within the same field. Different handling strategies are to be applied in order to exploit these different levels of variation.

4.2.1. Bulk variation

The geographical and between-field variations are easily utilised on the bulk level by handling cereal batches separately at the grain elevators. This handling has been greatly upgraded by the development of fast near infrared applications to grain analysis in recent decades. One example is protein determination, where tedious Kjeldahl analyses was first substituted by NIR analyses on flour, and at the present by NIT analysis on whole grains, which is performed on a batch before it is unloaded at the grain elevator. Loads with different qualities can thus be handled separately and collection can be optimised with regard to its further use.

Quality variations within the same field may be regarded in two different ways. Either one aims at reducing the quality variation across the field by site-specific precision agriculture, and thereby producing an improved and more homogenous grain lot from the field. Alternatively, one might utilise this quality variation by separating batches either directly in the field (selective harvest), or afterwards by sorting the bulked grains in different quality fractions.

Several studies have shown that crop quality within a field is highly variable (Mulla *et al.*, 1992; Thylén *et al.*, 1999). This was also seen in a preliminary investigation on Danish wheat (Jørgensen and Nielsen, 1999) where site-specific samples were taken within a 10 ha field (in Jutland) from the harvests of 1997 and 1998. In both years the crop was Terra wheat and 45 and 108 samples, respectively, were taken from the combine harvester throughout the field where site-specific fertilizer had been applied according to the yield of the preceding year and measurements of the crop requirements during the growing season. The range in protein content was 11.7 – 17.6 % and 9.6 – 13.6 % in the two years, respectively, showing that part of the field represents good bread wheat, while other parts of the field represent wheat for starch-requiring purposes, e.g. feed.

In precision agriculture the introduction of Global Positioning Systems (GPS) has made it possible to monitor within-field variability (Stafford, 2000). The idea is to monitor the fertility levels of the fields by soil and crop analysis, then transpose the data into fertilizer application equipment and thereby apply more fertilizer at field sites where more is needed, and vice versa. Initially, the main purpose was to improve yield, but more emphasis has been put on the use of this site-specific cultivation to improve quality. The above results on Terra wheat, together with other reports (Jørgensen and Jørgensen, 2001; Mulla *et al.*, 1992), point out that even though the fertilizer has been applied according to site-specific soil fertility and crop requirements, variations throughout the field are still seen. The existing variation, with or without attempting to reducing it, might not necessarily be seen as a problem, but could be considered an advantage. The question is how to sort the grains into different quality fractions?

Today, the GPS and yield meters on the combine harvesters are being extended with on-line protein sensors based on near infrared spectroscopy (Anonymous, 1999; Thylén and Algerbo, 2001). The protein content can hereby be assessed throughout the field and then assist in a more site-specific estimation of the nitrogen balance throughout the field (applied versus removed as estimated by grain yield in conjunction with protein content). Furthermore, already during harvest it would be possible to grade the grains into two or more fractions of different qualities. This selective harvest, of course, requires that spectroscopy-based bulk grading equipment is installed in the combine harvesters, and that the grain fractions are handled separately both in the combine harvester and in the subsequent handling. These are probable reasons why this approach has not yet been fully exploited.

4.2.2. Single seed variation

Grain quality variation is also seen within a site-specific position in the field, even within single seeds in the same plant (Home *et al.*, 1997; Angelino *et al.*, 1997). This grain quality variation cannot be utilised by applying selective harvest, since grains of different qualities are harvested together within the operation width of modern combine harvesters.

In order to utilise this single seed variation, sorting of the grains after harvesting is needed. It is important to emphasize that sorting is not the same as cleaning. Sorting is separation of grains due to their inherent properties, while cleaning is the removal of shrivelled kernels, broken kernels, etc. These waste kernels are easily removed by conventional cleaning machines and will not be discussed further.

There are two main potentials in sorting grains. Firstly, a grain lot with an average quality might be fractionated into two or more fractions with new bulk qualities, which might then be used for different purposes. Secondly, the fractionated grain lots will exhibit less variation (increased homogeneity) among the single kernels, which is of interest in both malting (Aastrup *et al.*, 1981) and milling (Ohm *et al.*, 1998). One example of an existing sorting machine is the gravity table, where each single seed is sorted according to its density. The new fractions then differ in kernel density, which as such might be interesting. However, the main potential in sorting for density is in sorting for other quality parameters, e.g. protein content, that might be correlated to density. In sorting and grading grains by size, form and density, the functional unit to be investigated is the single seed. Fast single seed quality analyses, ideally non-destructive, are therefore most important for an increased understanding of the variation in quality of the single seeds in a seed lot. This will provide an evaluation tool for the sorting potential and performance, and thereby be able to optimise the choice of variety, grading technique and end use.

The development of non-destructive screening methods for single seed protein, vitreousness, density and hardness index for single kernels of wheat is presented in Paper 6. A single kernel procedure involving image analysis, near infrared transmittance (NIT) spectroscopy, laboratory density determination, Single Kernel Characterization System (SKCS) and Kjeldahl protein determination on the crushed single kernels (the output grist of the SKCS) was applied.

Single kernel NIT spectroscopy showed excellent ability to determine protein content as shown by the predicted versus measured plots in Figure 4-2 for both calibration and test set with relative prediction errors of RE = 5.6 and RE = 4.8 for the calibration and test set, respectively. These results are comparable to earlier reported results (Delwiche, 1995; Delwiche, 1998).

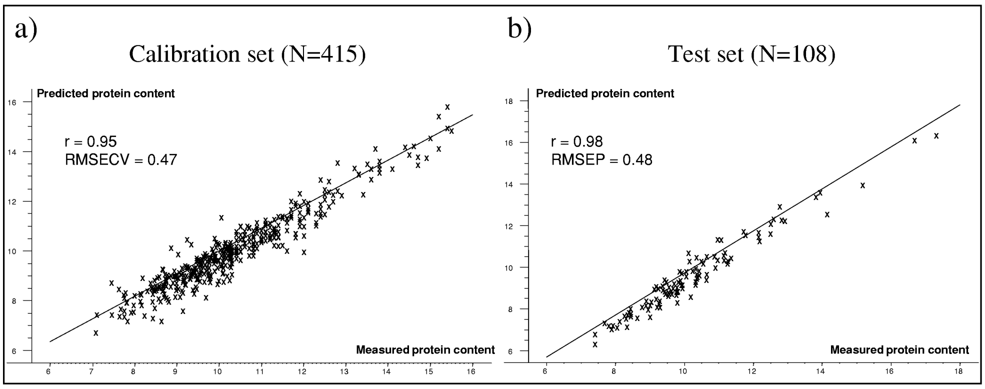


Figure 4-2. Predicted versus measured plot of a regression model for single seed protein using scatter-corrected NIT spectra for (a) the calibration set and (b) the subsequent prediction of test-set kernels

NIT spectroscopy has also been reported useful for determination of wheat hardness in bulk samples (Williams, 1991), and Delwiche (1993) reported on the use of single kernel NIT measurements for hardness determination. When calibrating single seed NIT spectra against bulk hardness data, he found that NIT spectra of single seeds had some ability to determine wheat hardness. In Paper 6, the SKCS hardness index as a true single kernel hardness reference in a NIT prediction model resulted in poor predictability. However, by applying an averaging approach (Paper 6), in which single seed replicate measurements are mathematically simulated, a very good NIT prediction model was achieved. This suggests that the single seed NIT spectra contain hardness information, but that a single seed hardness reference method with higher accuracy than the one performed currently by the SKCS instrument is needed in order to achieve a good NIT prediction model for single kernel hardness.

As reviewed in Chapter 2 and Paper 6, a range of successful single seed NIR and NIT measurements have been published during the recent years, rendering the method a potential tool for homogeneity analysis. The Infratec 1255 single seed instrument employed in Paper 6 provides excellent single seed protein data that are easier obtained than by the traditional Kjeldahl method, but the single seed handling is still not automated, and the measurements are quite time-consuming when analysing high number of kernels. Prior to any practical use of single seed near infrared spectroscopy as a homogeneity tool, it is necessary that the measurements are automated. This is done in the new combined SKCS-NIR instrument from Perten Instruments (Dowell *et al.*, 1999; Delwiche and Hruschka, 2000). When applied automatically, near infrared spectroscopy on single seeds, alone or in combination with other automated non-destructive techniques, has great potential as a routine homogeneity analysis for raw grain evaluation as well as for performance testing in the sorting industry. This need not be limited only to protein and hardness, but can also be utilised for other quality parameters in cereals, where the method currently used is limited to analysis on bulk samples.

As tools become more readily available for monitoring single seed quality, the question is how we apply these monitoring capabilities in order to utilise the normal single seed variability in the best way. Let us assume that one aims at sorting wheat grains for protein content by using density sorting on gravity tables, assuming there is a correlation between protein and density. The combinatory single seed approach in Paper 6 allowed us to explore the link between protein content and other single kernel characteristics on the exact same single seed i.e. on the exact same functional unit to be sorted. In Paper 6 a correlation (albeit on a limited number of kernels) between protein and density was found to be r=0.65. This relatively low correlation indicates that an indirect sorting for protein through sorting for density may not be sufficient. This agrees with earlier findings (Munck and Nielsen, 1998), where a two-tonne batch of Terra wheat was graded according to density on a full-scale density table (Cimbria Heid, Vienna). Apart from a fraction containing abnormal kernels (shrivelled), the different density fractions did not significantly differ in protein content (Munck and Nielsen, 1998). It should be noted that the sorting was only based on one variety, but laboratory testing of 36 Danish varieties using floatation separation of kernels based on kernel density (Hallgren and Murty, 1983; Munck, 1989) showed only minor differences in protein content between high- and low-density kernels (Munck and Nielsen, 1998).

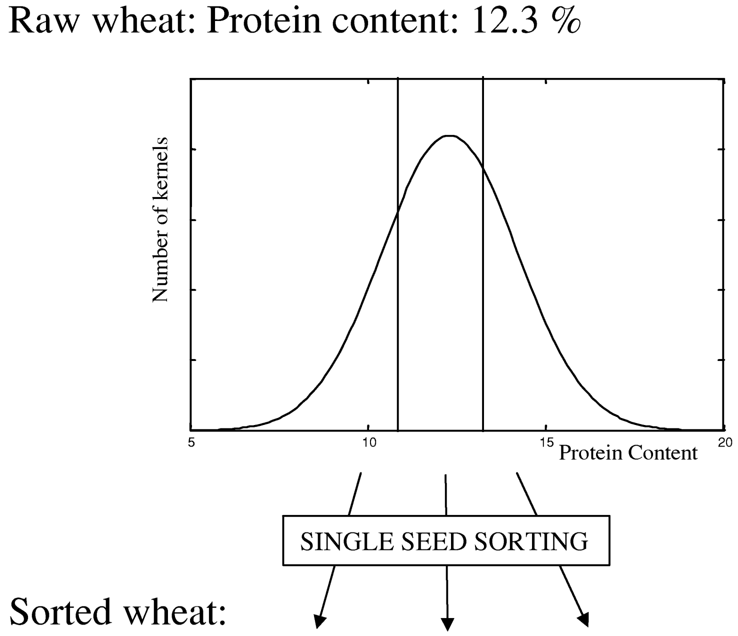
Thus on one hand, limitations are seen in the use of traditional sorting techniques for quality sorting of cereals, while on the other hand, encouraging quality determinations (e.g. for protein) using near infrared spectroscopy on single seeds have been demonstrated, as reviewed previously.

Methods exist by which single seeds are sorted according to the reflectance of a few filter-based wavelengths (max. three) chosen within the range of visible and infrared light. These methods are normally used to remove impurities or discoloured, infected or damaged kernels. In 1996, Brimrose Corp. (www.brimrose.com) introduced a laboratory instrument (Seed Meister) for single seed sorting of seeds based on full NIR spectra to be used in plant breeding. The spectra are acquired using the scanning technique of Acousto-Optic Tunable Filters (AOTF), and the setup records, predicts and sorts 30-40 kernels per minute (Hill, 2002).

In a recent patent application (Löfqvist and Nielsen, 2002), based on an idea introduced by BoMill AB, spectroscopic single seed assessment is used in a bulk sorting device for grains. The invention involves:

- distributing each single seed in a sorting device
- exposing each single seed to energy emitted from a light source
- ultra-fast recording of a multivariate spectrum from each seed
- ultra-fast prediction or classification based on chemometrics
- sorting of single seeds based on the prediction/classification result

A preliminary trial based on a laboratory-scale version of the invention has been conducted. A sub-sample of Capo wheat was drawn from a commercial silo.



Sorted wheat:

Protein Content (%)	10.2	12.0	14.4
Grain yield (%)	26	38	36

Figure 4-3. Results from a laboratory-scale sorting of wheat into three fractions of different protein content

This sub-sample showed a considerable single seed variation (Figure 4-3, upper histogram). The sorting device was adjusted to give three fractions; one below 11 % protein, one from 11 to 13 % protein and one above 13 % protein content (shown as vertical lines in the histogram in Figure 4-3). As seen from the included table, this wheat can be sorted into three fractions (nearly equal in amount) with considerable differences in protein content and therefore suitable for different purposes.

4.3. Monitoring of cereal processing using near infrared spectroscopy

The preceding sections in this chapter have dealt with analysis, handling and utilisation of the raw grains, i.e. the raw material in industrial processing of cereals. However, modern processing of cereals also calls for analysis methods of the intermediate materials and the final products in order to ensure that the products are within the quality specification of the end-user. On-line/at-line installations of NIR spectrometers can provide quality information during the processing, so that necessary adjustments can be made in time. In the following, the main focus will be on milling of wheat, although NIR spectroscopy is equally potential for monitoring the malting process, as shown by several reports on the development of at-line NIR applications used for monitoring the malting process (Garden and Freeman, 1998; Allosio *et al.*, 1997).

Industrial dry-milling of wheat consists of a complex procedure of consecutive steps of grinding and sifting. The aim is to obtain a high yield of endosperm flour without contamination of bran particles. The texture of the raw grains strongly influences the ease of processing the wheat; however, during the milling, adjustments of the different milling devices can be made by the miller in order to improve the yield and purity of the flour. In addition to the purity, the particle size distributions as well as other chemical parameters (moisture content, protein content, fibre content etc.) of the different flour outlets are important parameters to be controlled during the milling process. The fact that NIR analyses are rapid and non-destructive and the spectra are fingerprints of the particle size and the chemical properties of the flour samples makes NIR spectroscopy an obvious tool for monitoring and process control in the milling industry. As reviewed by Osborne *et al.* (1993), NIR has been widely used for determination of chemical compounds in wheat flour. The link between NIR spectra and particle size of flours has been extensively investigated (Chapelle *et al.*, 1989; Obsorne *et al.*, 1981) and NIR has recently been evaluated as a granulation sensor for first break flour (Pasikatan *et al.*, 2002).

Paper 7 is a chemometric study in which the relationship between NIR spectra, particle size distributions and chemical properties of flour samples was investigated simultaneously by using a qualitative multi-way data analysis called “Analysis of Common Dimensions and Specific Weights” (COMDIM) (Quannari *et al.*, 2000; Paper 7). Six wheat flour streams from a full-scale mill were sampled together with the final flour (proportional mixture of the six first flours). Each of these seven samples was fractionated into six sub-samples according to particle size (sieving). The 42 samples were then recorded by NIR (NIRSystems 6500, Foss NIRSystems) and analysed for particle size distributions and chemical composition (moisture, starch, ash, protein and damaged starch).

The COMDIM was applied on three data tables representing the NIR spectra, the particle size data, and the chemical compositional data of the wheat flour samples. The output from this analysis that conceptually can be seen as a PCA across several data tables has been useful in interpreting the relationship between the structural and chemical composition of flours. In this way, the importance of the information from the different flour measurements has been assessed and interpreted in a more straightforward manner than by doing PCA on three data tables separately. The spectral, particle size and chemical loadings of the underlying dimensions were interpretable and showed patterns in agreement with prior knowledge regarding characteristics of wheat flour. The exact same data set has been used for multi-block analysis (van den Berg, 2001) who by applying multi-block PCA and multi-block PLSR (Westerhuis *et al.*, 1998) achieved corresponding results.

The multi-way data analysis applied here represents a useful exploratory tool when NIR spectroscopy is applied for quality control in a flour mill. Normally, acquired NIR spectra are used in a quantitative way, as in the prediction of chemical constituents using PLSR. Instead of doing so, it is suggested to perform a "qualitative calibration" where the flour samples are ranked according to their NIR score values, or a combination of several scores, for example if the miller prefers samples having a certain value of the first score and a certain value of the second score, for example representing a gritty flour with low starch damage. Such a qualitative calibration can, of course, also be done by a classical PCA on the NIR spectra, but the advantage of this multi-way approach is that the variation in both the particle size and chemical data is used simultaneously to guide the decomposition of the spectral data. This mimics the skill and "fingerspitzengefühl" of the miller, and can be used by him for further optimisation of different quality parameters or for an automatisation of the milling process.

5. Concluding remarks and perspectives

This thesis describes the development and application of fast multivariate methods combined with chemometrics for the task of monitoring, handling and utilisation of quality variations of barley and wheat. The potential in this approach is demonstrated in the presented examples which cover applications for use within plant breeding, handling and sorting, and cereal processing, all of which may assist in improving the end products from the barley and wheat industry.

The investigations have been limited to barley and wheat quality; however, the methods might equally well be used for other cereals and other plant products in which a comprehensive quality monitoring is needed. Most of the included investigations have been limited to a relatively low number of samples and the presented results are only indications of potential applications. The results are to be confirmed and even improved through the building of global models including more samples covering a broader range. In order to do this, a global sample/data library is to be developed in the same way as the NIT Network has been built to give robust global calibrations using NIT spectroscopy and neural network on huge data bases.

The fast multivariate instruments applied, including NIT and NIR spectroscopy, image analysis and hardness analysis, are all used as multivariate fingerprints of the samples either in bulk or on single seed. In combination with PLSR these multivariate fingerprints have been successfully used in the development of fast prediction models for various traditional quality parameters in barley and wheat (Paper 2, 3 and 6). This strategy could be seen as the way the instrument suppliers deliver new instruments today, including a calibration for identified quality parameters in a “blackbox - press the button and read the result” instrument. For the purpose of simply substituting slow and costly laboratory analyses, this is feasible.

In parallel to this, a more qualitative exploratory strategy, typically using PCA, has been applied in several of the investigations and found to be a very powerful tool, for instance, in order to explore the underlying factors of morphological data and hardness data of barley kernels for a malting barley characterisation (Paper 2) or for exploring near infrared spectra of mutants in barley (Paper 5). Facilitated by graphic displays and without too many restrictive *a priori* assumptions, this can lead to new and unexpected correlations and hypotheses. Thus, for internal use in both the cereal industry and in cereal research, a more exploratory software in cereal quality instruments would allow the user (breeder, malster, brewer, miller, baker, researcher, etc.) to utilise the multivariate output of the instruments in a more exploratory way. In this way, the cereal knowledge, the multivariate measurements and the chemometric evaluations are brought together – a powerful combination which might be introduced as a new chemometric sub-discipline: "CereaMetrics".

Traditional analyses determined either according to approved methods in AACC, EBC etc. or as predictions by multivariate methods are used throughout the cereal industry as the "quality language". However, each of these quality analyses reflects only a univariate part of a complex multivariate functional quality of a barley or wheat sample for industrial processing. Thus, multivariate methods are used for predicting of univariate quality parameters, several of which are subsequently combined by the end-user in order to evaluate a complex multivariate functional quality. This calls for a more multivariate approach, as demonstrated by the exploration of malting data in Paper 1 using PCA and in Paper 4 where fuzzy logic is successfully applied for the calculation of an overall quality index. Paper 4 moreover indicates that it might be possible to go even a step further and use multivariate data (NIT spectra) for direct prediction of the multivariate based end quality, validated by the experience of a malting expert. The multi-way approach in Paper 7 is also a step in the direction of a multivariate monitoring of the "fingerspitzengefühl" of the miller. Among the long-term perspectives of these ideas may be that fast multivariate spectra, including hundreds or thousands of variables, could be used as quality fingerprints of the samples instead of, for example, protein content, starch content etc. By defining spectra of good and bad cereal samples for a given purpose, chemometric tools can be used to decide whether or not a given sample is suitable for a given purpose. Thus, within the limits of the multivariate sensors, fast instrumental methods in combination with chemometrics may be the new "quality language" to be used in the cereal industry.

Thanks to new instrumentation and larger computers, fast single seed analysis has become possible. By combining several single seed methods (Paper 6), the relationships among several parameters can be studied on the exact same kernel and fast prediction models can be developed.

From a cereal processing point-of-view, the preliminary single seed results showing that it is possible to sort single seeds in bulk based on quality as predicted by spectroscopic measurements is of great practical interest. A full-scale sorting device based on this principle will make it possible to fractionate average wheat quality into fractions of different qualities. This could, for example, be sorting into a low-protein fraction for feed and a high-protein fraction for bread production or into fractions of “acceptable overall malting quality” and “unacceptable overall malting quality”. In this way, many of the environmental quality differences may be utilised (and not regarded as problematic), even if the grains from a whole field or several fields are harvested and mixed in one batch as is the common practice today. This approach could, of course, also be applied to other crops such as corn, soy or coffee beans.

6. References

Aastrup, S., Gibbons, G. C. and Munck, L. (1981). A rapid method for estimating the degree of modification in barley malt by measurements of cell wall breakdown. *Carlsberg Research Communication* 46, 77-86.

Allison, M.J. (1989). Areas of absorption relating to malt extract value in modified near infra-red spectra of barley flour. *Journal of the Institute of Brewing* 95, 283-286.

Allosio, N., Boivin, P., Bertrand, D. and Courcoux, P. (1997). Characterisation of barley transformation into malt by three-way factor analysis of near infrared spectra. *Journal of Near Infrared Spectroscopy* 5, 157-166.

Angelino, S.A.G.F., van Laarhoven, H.P.M., van Westerop, J.J.M., Broekhuijse, B.M. and Mocking, H.C.M. (1997). Total nitrogen content in single kernel of malting barley samples. *Journal of the Institute of Brewing* 103, 41-46.

Anonymous. (1983). Approved methods of the American Association of Cereal Chemists. Eighth edition, Volume I and II. American Association of Cereal Chemists, Inc., Minnesota, USA.

Anonymous. (1987). Analysis by the European Brewery Convention. 4.edition. Brauerei- und Getränke-Rundschau, Zurich.

Anonymous. (1999). Case Textron System. *Spectroscopy* 14 (10), 54.

Barnes, R.J., Dhanoa, M.S. and Lister, S.J. (1989). Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy* 43 (5), 772-777.

Berg, F.V.D. (2001). Multi-block PLSR models in Food technology. In: PLS and related methods; Proceedings of the PLS'01 International Symposium pp. 385-394; Eds: Vinzi, V.E., Lauro, C., Morineau, A. and Tenenhaus, M.

Berman, M., Bason, M. L., Ellison, F., Peden, G. and C. W. Wrigley. (1996). Image Analysis of Whole Grains to Screen for Flour-Milling Yield in Wheat Breeding. *Cereal Chemistry* 73, 323-327.

Büchmann, N.B., Josefsson, H. and Cowe, I. A. (2001). Performance of European Artificial Neural Network (ANN) Calibrations for Moisture and Protein in Cereals Using the Danish Near-Infrared Transmission (NIT) Network. *Cereal Chemistry* 78 (5), 572-577.

Campbell, M.R., Sykes, J. and Glover, D.V. (2000). Classification of Single- and Double-Mutant Corn Endosperm Genotypes by Near-Infrared Transmittance Spectroscopy. *Cereal Chemistry* 77 (6) 774-778.

Chapelle, V., Melcion, J.P., Robert, P. and Bertrand, D. (1989). Application of Near Infrared spectroscopy to particle size analysis of a pea flour. *Sciences des aliments* 9, 387-404.

Chtioui, Y., Bertrand, D., Dattée, Y. and Devaux, M-F. (1996). Identification of Seeds by Colour Imaging: Comparison of Discriminant Analysis and Artificial Neural Network. *Journal of the Science of Food and Agriculture* 71, 433-441.

Day, M. (1999). Barley and malt analysis using near infrared spectroscopy – the U.K. experience. *The Brewer*, December 1999, 612-616.

De Noord, O.E. (1994). The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems* 23, 65-70.

Delwiche, S. R. (1998). Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. *Journal of Cereal Science* 27, 241-254.

Delwiche, S. R. and Hruschka, W. R. (2000). Protein content of bulk wheat from near-infrared reflectance of individual kernels. *Cereal Chemistry* 77(1), 86-89.

Delwiche, S. R. and Massie, D. R. (1996). Classification of wheat by visible and near-infrared reflectace from single kernels. *Cereal Chemistry* 73(3), 399-405.

Delwiche, S. R., (1995). Single wheat kernel analysis by near-infrared transmittance: protein content. *Cereal Chemistry* 72(1), 11-16.

Delwiche, S.R. and Graybosch, R.A. (2002). Identification of Waxy Wheat by Near-infrared Reflectance Spectroscopy. *Journal of Cereal Science* 35, 29-38.

Delwiche, S.R., Graybosch, R.A. and Peterson, C.J. (1999). Identification of Wheat Lines Possessing the 1AL.1RS or 1BL.1RS Wheat-Rye Translocation by Near-Infrared Reflectance Spectroscopy. *Cereal Chemistry* 76 (2), 255-260.

Delwiche,S. (1993). Measurement of single-kernel wheat hardness using near-infrared transmittance. *Trans. ASAE* 36(5), 1431-1437.

Dowell, F. E. (2000). Differentiating vitreous and nonvitreous durum wheat kernels by using near-infrared spectroscopy. *Cereal Chemistry* 77(2), 155-158.

Dowell, F.E., Ram, M.S. and Seitz, L.M. (1999). Predicting Scab, Vomitoxin, and Ergosterol in Single Wheat Kernels Using Near-Infrared Spectroscopy. *Cereal Chemistry* 76 (4), 573-576.

Faccioli, P., Cavallero, A., Stanca, A.M., Fornasier, F. and Odoardi, M. (2000). A rapid way of evaluating barley grain for malting quality and as functional food. In Near Infrared Spectroscopy: Proceedings of the 9th International Conference, Ed. by A.M.C. Davies and R. Giangiacomo. NIR Publications, Chichester, UK, pp. 689-691 (2000).

Fearn, T. (2001). Chemometrics for near infrared spectroscopy: past present and future. *Spectroscopy Europe* 13 (2), 10-14.

Garden, S.W. and Freeman, P.L. (1998). Applications of Near-Infrared Spectroscopy in Malting: Calibrations for Analysis of Green Malt. *Journal of the American Society of Brewing Chemists* 56 (4), 159-163.

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

Hallgren, L. and Murty, D. (1983). A screening Test for Grain Hardness in Sorghum Employing Density Grading in Sodium Nitrate Solution. *Journal of Cereal Science* 1, 265-274.

Halsey, S.A. (1987). Analysis of whole barley kernels using near infrared reflectance spectroscopy. *Journal of the Institute of Brewing* 93, 461-464.

Helland, I.S,. Næs, T. and Isaksson, T. (1995). Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* 29, 233-241.

Henry, R.J. (1985). Evaluation of barley and malt quality using near-infrared reflectance techniques. *Journal of the Institute of Brewing* 91, 393-396.

Henry, R.J. (1985a). Use of Scanning Near-infrared Reflectance Spectrophotometer for Assessment of the Malting Potential of Barley. *Journal of the Science of Food and Agriculture* 36, 249-254.

Hill, J. (2002). Personal communication. Brimrose Corp., (www.brimrose.com).

Home, S., Wilhelmson, A., Tammisola, J. and Husman, J. (1997). Natural Variation Among Barley Kernels. *Journal of the American Society of Brewing Chemists* 55 (2), 47-51.

Höskuldsson, A. (1994). The H-principle: new ideas, algorithms and methods in applied mathematics and statistics. *Chemometrics and Intelligent Laboratory Systems* 23, 1-28.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417-441.

Isaksson, T. and Kowalski, B. (1993). Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products. *Applied Spectroscopy* 47, 702-709.

Jacobsen, T. (1980). New aspects of chemical analysis: Chemometrics – Variations in trace elements in malt as shown by factor analysis. *Brygmesteren* 6, 149-158.

Jørgensen, J.R., Jørgensen, R.N. (2001). Impact on grain quality when nitrogen is 'sensor applied' by the 'hydro precise system'. Proceedings of the Third European

Conference on Precision Agriculture (ed. G. Grenier and S. Blackmore), Vol. 2, 929-934.

Jørgensen, J. R. and Nielsen, J. Pram. 1999. Utilization of site specific variation of grain characteristics within wheat crops. Abstract Proceeding of the Second European Conference on Precision Agriculture '99, Odense, Denmark.

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

Lindholm, P. (2002). Personal communication. Foss Tecator (www.foss.dk).

Löfqvist, B. and Nielsen, J. P. (2002). A method of sorting objects comprising organic material. International patent application (Patent Cooperation Treaty).

Lonkhuijsen, H. J. v., Douma, A. C. and Angelino, S. A. G. F. (1998). Evaluation of a Malting Barley Quality Assessment System. *Journal of the American Society of Brewing Chemists* 56 (1) 7-11.

Madre, M. (2002). Malting barley in France: recent trends and current situation. *Bios* 5, 24-31.

Martens H and Næs T. (1989). Multivariate Calibration. Wiley, New York.

Martens H. and Jensen S.A. (1983). Partial Least Squares Regression: A new two-stage NIR calibration method. Proceedings of the VIIth World Cereal and Bread Congress, Prague, 1982, pp. 607-647. Eds. Holas, J. and Kratochvil, J., Elsevier Science Publishers, Amsterdam.

Martens, H. and Martens, M. (2000). Modified Jack-knife estimation of parameters uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference* 11, 5-16.

Martens, H. and Stark, E. (1991). Extended multiplicative signal correction and spectral interference substraction: new preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical & Biomedical Analysis* 9, 625-635.

Martens, H., Jensen, S. A., Geladi, P. (1983). Nordic Symposium on Applied Statistics, Stokkand Forlag Publ.pp. 208-34.

Martens, H., Nielsen, J.P. and Engelsen, S.B. (2002). Light scattering and light absorbance separated by Extended Multiplicative Signal Correction (EMSC). Application to NIT analysis of powder mixtures. *Analytical Chemistry*, accepted (Appendix 2).

Martin, C., Rousser, R. and Brabec, D. 1993. Development of a single-kernel wheat characterization system. *Trans. ASAE* 36:1399-1404.

Meurens, M. and Yan, S.H. (2002). Applications of Vibrational Spectroscopy in Brewing. In: Handbook of Vibrational Spectroscopy; Applications in Life, Pharmaceutical and Natural Science; Eds: Chalmers, J. M. and Griffiths, P.R. John Wiley and Sons, LTD, Chicester, UK.

Molina-Cano, J.L. (1987). The EBC Barley and Malt Committee Index for the Evaluation of Malting Quality in Barley and its Use in Breedning. *Plant breeding* 98, 249-256.

Molina-Cano, J.L., Madsen, B., Atherton, M. J., Drost, B.W., Larsen, J., Schildbach, R., Simiand, J.P. and Voglar, K. (1986). A Statistical Index for the Overall Evaluation of Malting and Brewing Quality in Barley. *Monatsschrift für Brauwissenschaft* 9, 328-335.

Mulla, D.J., Bhatti, A.U., Hammond, M.W. and Benson. J.A. (1992). A comparison of winter wheat yield and quality under uniform versus spartially variable fertilizer management. *Agriculture, Ecosystems and Environment* 38, 301-311.

Munck, L. (1989). Supporting wheat plant breeding and technology with new analytical tools – some experiences and perspectives. In: Proceedings form ICC '89 Wheat end-use properties. Wheat and Flour Characterization for Specific End-Uses, Lathi, Finland. Ed. by Solovaara, H.

Munck, L. (1991). Quality criteria in the production chain from malting barley to beer. *Ferment* 4, 235-241.

Munck, L. (1992). The Case of High-lysine Barley Breeding. In: Barley Genetics, Biochemistry, Molecular Biology and Biotechnology; Ed: Shewry, P.R. C.A.B. International, Wallingford, Oxon, UK.

Munck, L. and Nielsen, J. Pram. (1998). Quality Control – Individual Progress Report. In: Midterm Report for the FAIR CT96-1105 "Cascade Refining of European Wheat for Production of High-Quality Products for the Paper Industry". Ed: Munck, L., Food Technology, Dept. of Dairy and Food Science, Copenhagen.

Nielsen, J.P. (2001). NIT prediction of the similarity to an "ideal" malt quality profile. Poster presented at the Annual meeting in the Danish Cereal Network, November 2001.

Nielsen, J.P. and Munck, L. (2000). Exploring Malting Barley Data using Chemometrics. In: Proceedings of the 8th International Barley Genetics Symposium, 22-27 October 2000, Adelaide, South Australia. Ed. by Louge, S., pp. 258-261.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L. and Engelsen, S.B. (2000) Interval partial least squares regression (iPLS): A comparative chemometric study with an example from near infrared spectroscopy. *Applied Spectroscopy* 50 (4) 413-419 (Appendix 1).

Ohm, J.B., Chung, O.K. and Deyoe, C.W. (1998). Single-Kernel Characteristics of Hard Winter Wheats in Relation to Milling and Baking Quality. *Cereal Chemistry* 75 (1), 156-161.

Osborne, B. G. and Douglas, S. (1981). Measurement of the Degree of Starch Damage by Near Infrared Reflectance Analysis. *Journal of the Science of Food and Agriculture* 32, 328-332.

Osborne, B., Fearn, T. and Hindle, P.H. (1993). Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman Scientific and Technical, Harlow, UK.

Osborne, B.G., Kotwal, Z., Blakeney, A.B., O'Brien, L., Shan, S. and Fearn, T. (1997). Application of the Single-Kernel Characterization System to Wheat Receiving Testing and Quality Prediction. *Cereal Chemistry* 74 (4)

Pasikatan, M.C., Steele, J.L., Haque, E., Spillman, C.K. and Miliken, G.A. (2002). Evaluation of a Near-Infrared Reflectance Spectrometer as a Granulation Sensor for First-Break Ground Wheat: Studies with Hard Red Winter Wheats. *Cereal Chemistry* 79 (1), 92-97.

Pearson K. (1901). On lines and planes of closest fit to systems of points in shape. *Philosophical Magazine* 2, 559-572.

Pedersen, D.K., Martens, H., Nielsen, J.P. and Engelsen, S.B. (2002). Near infrared absorption and scattering separated by Extended Inverted Signal Correction (EISC). Analysis of NIT spectra of single wheat seeds. *Applied Spectroscopy* 56 (9) 1206-1214, (Appendix 3).

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

Roumeliotis, S., Logue, S.J., Jefferies, S.P. and Barr, A.R. (2000). The development and implementation of near infrared calibrations for predicting malting quality in barley. In Near Infrared Spectroscopy: Proceedings of the 9th International Conference, Ed. by A.M.C. Davies and R. Giangiacomo. NIR Publications, Chichester, UK, pp. 673-678.

Ruan, R., Ning, S., Song, A., Ning, A., Jones, R. and Chen. P. (1998). Estimation of *Fusarium Scab* in Wheat Using Machine Vision and a Neural Network. *Cereal Chemistry* 75 (4) 455-459.

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

Scotter, C. (1990). Use of near infrared spectroscopy in the food industry with particular reference to its application to on/in-line food processes. *Food Control*, July 1990, 142-149.

Siebert, K. J. (2001). Chemometrics in Brewing – A Review. *Journal of the American Society of Brewing Chemists* 59 (4) 147-156.

Sinnaeveg, G., Dardenne, P. and Biston, R. (1994). Potentialites de la spectrometrie dans le proche infrarouge (SPIR) pour l'evaluation de la qualite brassicole des orges et de la qualite des malts. *Cerevisia and Biotechnology* 19 (2), 21-25.

Stafford, J.V. (2000). Implementing Precision Agriculture in the 21st Century. *Journal of Agricultural Engineering Research* 76, 267-275.

Szczodrak, J., Czuchajuwska, Z. and Pomeranz, Y. (1992). Characterisation and Estimation of Barley Polysaccharides by Near-Infrared Spectroscopy. II. Estimation of Total ß-D-glucans. *Cereal Chemistry* 69 (4), 419-423.

Thylén, L. and Algerbo, P.A. (2001). Development of a protein sensor for combine harvesters. In: Procceding of the Third European Conference of the European Federation for Information Technology in Agriculture, Food and Environment (EFITA), Ed. by Steffe, J.

Thylén, L., Algerbo, P.A., Pettersson, C.G. (1999). Grain quality variations within fields of malting barley. In: Precision agriculture ´99; Ed: Stafford J.V. Sheffield Academic Press, pp 287-296.

Wang, D., Dowell, F.E. and Dempster, R. (2002). Determining Vitreous Subclasses of Hard Red Spring Wheat Using Visible/Near-Infrared Spectroscopy. *Cereal Chemistry* 79 (3), 418-422.

Wang, D., Dowell, F.E. and Lacey, R.E. (1999). Predicting the Number of Dominant R Alleles in Single Wheat Kernels Using Visible and Near-Infrared Reflectance Spectra. *Cereal Chemistry* 76 (1), 6-8

Westad, F. and Martens, H. (2000). Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression. *Journal of Near Infrared Spectroscopy* 8, 117-124.

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

Wiliams, P. (1991). Prediction of Wheat Kernel Texture in Whole Grains by Near-Infrared Transmittance. *Cereal Chemistry* 68 (1), 112-114.

Wiliams, P. and Sobering, D.C. (1993). Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. *Journal of Near Infrared Spectroscopy* 1, 25-32.

Williams, P. (2002). Near-infrared Spectroscopy of Cereals. In: Handbook of Vibrational Spectroscopy; Applications in Life, Pharmaceutical and Natural

Science; Eds: Chalmers, J. M. and Griffiths, P.R. John Wiley and Sons, LTD, Chicester, UK.

Williams, P. and Norris, K. (1987). Near Infrared Technology in the Agricultural and Food Industries. American Association of Cereal Chemists, Minnesota, USA.

Wold, S., K. Esbensen, and P. Geladi. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 37-52.

Wold, S., Martens, H. and Wold, H. (1983). The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. Proc. Conf. Matrix Pencils, (A. Ruhe and B. Kågström, eds.), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 286-293

Zayas, I., Lai, F. S. and Pomeranz, Y. (1986). Discrimination Between Wheat Classes and Varieties by Image Analysis. *Cereal Chemistry* 63, 52-56.

Zayas, I.Y., Martin, C.R., Steele, J.L. and Katsevich, A. (1996). Wheat Classification Using Image Analysis and Crush-force Parameters. *Trans. ASAE* 39 (6), 2199-2204.

Paper 1

Evaluation of malting barley quality using exploratory data analysis. I. Extraction of information from micro-malting data of spring and winter barley

Jesper Pram Nielsen and Lars Munck

Abstract

This paper presents an exploratory multivariate approach for analysis of malting barley quality data. By using Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR), complex malting quality data are combined into functional factors which are used in a malting barley quality characterisation. Fifty barley samples were used in this investigation, representing 15 spring barley and 10 winter barley varieties grown at two locations in Denmark. The samples were micro-malted and mashed, and analysed for 13 quality parameters according to the official methods of the European Brewery Convention. These data were combined and reduced into a few latent (functional) factors using Principal Component Analysis (PCA) by which it is demonstrated that the modification of ß-glucan plays a major role in both spring and winter barleys. Additionally, the spring barley and winter barley samples display different covariate latent structures, mainly in the nitrogen and diastatic power patterns. It is furthermore shown that graphic display as facilitated by exploratory data analysis can be utilized in order to evaluate genotype-environmental interactions by considering the position and movements of the individual objects (here genotypes) in the score plots. Thus, in contrast to the classical analysis of variance, the samples can be individually evaluated and the corresponding loadings can be used to validate the genetic and environmental effect of a given sample in a quality perspective.

Several of the investigated malting quality parameters are be highly intercorrelated. This fact is utilized by applying PLSR on barley and malt data for the prediction of wort quality in order to exclude the mashing step. This approach was successful for the modification-dependent wort parameters extract, ß-glucan in wort and viscosity.

Introduction

An increasing number of quality criteria are used in the evaluation of malting barley. A complete quality analysis includes an elaborate and expensive simulation of the malting and mashing steps in micro-scale, and could involve 10-15 physical and chemical parameters from analysis of the raw barley, the intermediate malt and the final wort. The quality of the wort should in this context reflect the demands of the brewer and the properties of the barley and malt are considered as indirect predictors of the wort quality. In the classical quality evaluation, each parameter from the raw barley, the malt and the wort data are carefully controlled to be within the limits of the specifications. However, the results of these different analyses are not independent. In fact, they form characteristic relationships, which by means of an exploratory Principal Component Analysis (PCA) and validation could be identified as functional factors to be utilized in a characterisation of malting barley quality.

The terms multivariate data analysis and chemometrics are often used synonymously. However, multivariate data analysis is broader and refers to multivariate data analysis applied in any field of science, while chemometrics is the application of multivariate mathematics to efficiently extract maximum useful information from chemical data. Multivariate data analysis facilitates a graphic displaying of the underlying latent factors (principal components) and the interface between the individual samples and variables. Multivariate data analysis allows for an exploratory data analytical strategy, in which complex data sets are explored with a minimum of pre-assumptions, by mathematically reducing their dimensionality into fundamental underlying factors followed by validation against prior knowledge. Thus, with less *a priori* knowledge, exploratory data analysis can be exploited to achieve a better understanding of complex data sets.

Back in 1971, Reiner[1] used factor analysis to reduce 49 parameters from barley, malt and beer into four significant factors representing water uptake, yield/protein/extract, kernel development and cytolytic modification, respectively. Jacobsen (1980)[2] showed that it was possible to distinguish growing location of the barley, growing year and micro-malting laboratory using Principal Component Analysis on trace elements in malt. Munck (1991)[3] demonstrated the use of Principal Component Analysis (PCA) in order to compress a complex malt quality data set into a few principal components interpretable in physical and chemical terms.

The purpose of this paper is to characterize a set of malting barley quality data from spring and winter barley grown at two different locations in Denmark by applying exploratory data analysis. We will demonstrate the use of graphically oriented exploratory tools for a holistic evaluation of the interaction between malting quality characteristics due to genotypes and environment. The varieties are potential malting barley varieties bred by members of the European Brewery Convention (EBC). The samples in this investigation show considerable variation in malting quality. The material is thus suitable for an initial study of the malting barley quality complex. In a subsequent paper[4] the same sample set is used for studying the use of barley kernel morphology and hardness as a screening method.

Experimental

Sample collection:

The barley samples originate from trials under EBC, harvested in 1995. Fifteen spring barley varieties and ten winter barley varieties were grown at two different locations in Denmark, Jutland and Zealand, providing 50 malting barley samples in total. These are listed in Table I. The barley has been cultivated according to common Danish practice, including common fertilizer and plant protection regimes. The barley grain samples were screened over a 2.5-mm sieve and the grains above 2.5 mm were subjected to micro-malting procedure.

Table I: List of the 15 spring barley and 10 winter barley varieties grown in Jutland and Zealand.

Number	Spring barley varieties	Number	Winter barley varieties
1	Alexis	16	Plaisant
2	Triumph	17	Angora
3	Nevada	18	Clarine
4	Cooper	19	Puffin
5	Caminant	20	Geneva
6	Miralix	21	Trasco
7	Texana	22	Fanfare
8	Trebon	23	Melanie
9	Cork	24	Rejane
10	Delibes	25	Sunrise
11	Polygena		
12	Mentor		
13	Mie		
14	Reggae		
15	Anni		

Malting quality analyses:

The screened barley grain samples were micro-malted on an “Automatic Micromalting System” (Phoenix Systems) involving 15 hours of steeping, 84 hours of air rest (germination) and 31 hours of kilning. The following 13 quality analyses were performed according to the official methods of the EBC[5]: ß-glucan in barley (BGIB) and in malt (BGIM), extract (EXTRACT), nitrogen in malt (NIM), Kolbach Index (KOLBACH), viscosity (VISCOS), friability (FRIAB), malt modification (MODIF), malt homogeneity (HOM), diastatic power (DP), wort colour (WORTCOL), nitrogen in wort (NIW) and ß-glucan in wort (BGIW).

Exploratory data analysis:

Principal Component Analysis (PCA)[6] and Partial Least Squares Regressions (PLSR)[7] were performed using The Unscrambler version 7.6 SR-1 (CAMO A/S, Norway). Since the methods employed are described in detail in the literature, only a brief description will be given below. The data was mean centred and weighted (auto scaling) by the standard deviation prior to the PCA and PLSR calculations in order to take into account that the measured variables were in different units.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA)[6] is probably the most fundamental chemometric algorithm. The PCA algorithm finds the main directions in a multidimensional data set by creating new orthogonal (i.e. independent) linear combinations (principal components) of the raw data. These linear combinations approximate the original data set in a least squares sense. PCA provides an approximation of the data matrix (malting quality profiles) in terms of the product of the two low dimensional matrices **T** (scores) and **P'** (loadings), where ' is transposed. These two matrices capture the systematic variation of the data matrix. The columns in **T** (scores) contain information about the samples and the rows in **P'** (loadings) contain information about the variables. The loadings are common to all samples, and the scores specify the amount (concentration) of the common loadings within each of the samples. Plots of the columns of **T** (score plots) provide a picture of the sample concentrations of the latent variables, while plots of the rows of **P'** (loading plots) depict the variable contribution to the latent variables.

Partial Least Squares Regression (PLSR)

Partial Least Squares Regression[7] is a predictive two-block regression method also based on latent variables and is applied to the simultaneous analysis of two data matrices. The purpose of the PLSR is to build a linear model between a desired

y characteristic (e.g. extract yield) from easily obtainable **X** data (e.g. barley and malt quality data). In PLSR, a multiple linear regression model is built between the significant scores (**T**) and the **y**. As compared to the PCA scores described above, the significant PLSR scores **T** are found in a slightly different way, taking into account the variation in **y** during the decomposition of **X**, i.e. the covariance between **X** and **y** is maximized[7]. The number of factors (components) to include is found by validation, preferably test set validation. In the case of small data sets, as in this investigation, cross-validation is the alternative way of validating the model number of components and estimating the predictive ability.

The root mean square error of cross-validation (RMSECV) in combination with the correlation coefficient (*r*) is used as a measure of how well a given cross-validated model performs. The RMSECV is denoted as follows:

$$\text{RMSECV} = \sqrt{\frac{\sum_{N}(y_{pred} - y_{ref})^2}{N}}$$

where y_{pred} is the predicted value using cross-validation, y_{ref} is the laboratory measured value, and N is the number of samples. The relative error (RE) in percent is calculated as:

$$\text{RE} = \left(\frac{\text{RMSECV}}{\text{value}_{\text{max}} - \text{value}_{\text{min}}}\right) * 100$$

where RMSECV is the cross-validated prediction error, the $\text{value}_{\text{max}}$ is the highest value and the $\text{value}_{\text{min}}$ is the lowest value of the **y** parameter in question.

The importance of the **X** variables in PLSR models can be evaluated using the modified Jack-knife validation proposed by Martens and Martens (2000)[8]. In this method the regression coefficients of all the sub-models in the cross validation are calculated and used to estimate the uncertainty of each variable regressor in the PLSR models. It is hereby possible to inspect each of the variables in **X** used in the PLSR with respect to both importance (magnitude) and uncertainty.

Results and Discussion

Exploring the malting barley quality data using PCA:

Results of the 13 malting barley quality analyses are given in Table II. The results are presented as averages and ranges between samples of the spring barley varieties grown in Zealand (●), spring barley varieties grown in Jutland (○), winter barley varieties grown in Zealand (■) and winter barley varieties grown in Jutland (□). The sample set shows considerable variation in malting quality. Good malting behaviour is normally characterised by low ß-glucan in barley, malt and wort, low viscosity, high extract content, high friability, high modification, high homogeneity and high diastatic power together with a low level of nitrogen in malt and medium levels of nitrogen in wort, Kolbach Index and wort colour.

From the averages and ranges in Table II, it is obvious that the winter varieties grown in Jutland differ from the three other classes in having an overall lower malting quality, especially in terms of higher viscosity, higher ß-glucan in malt, higher ß-glucan in wort, lower modification and friability. The winter barley varieties grown in Zealand seem to be superior to the winter barley varieties grown in Jutland and are almost as good as the spring barleys. It is also noteworthy that the winter barley varieties grown in Jutland have considerably higher levels of diastatic power, nitrogen in malt and wort than the other three sub-materials.

Table II. Mean and ranges (in parentheses) of samples of the malting barley quality data of the four sub materials.

	ID	Spring barley Zealand (●) N=15	Spring barley Jutland (○) N=15	Winter barley Zealand (■) N=10	Winter barley Jutland (□) N=10
1	ß-glucan in barley (%)	3.87 (3.14 - 4.60)	4.05 (3.49 - 4.98)	3.09 (2.77 - 3.42)	3.60 (3.20 - 3.92)
2	ß-glucan in malt (%)	0.48 (0.21 - 1.52)	0.65 (0.26 - 1.79)	0.50 (0.21 - 1.10)	1.24 (1.00 - 1.80)
3	Extract yield (%)	82.8 (81.7 - 83.9)	81.6 (80.9 - 82.4)	81.8 (80.6 - 82.7)	78.9 (77.1 - 80.5)
4	Nitrogen in malt (%)	1.47 (1.35 - 1.55)	1.60 (1.49 - 1.71)	1.46 (1.33 - 1.74)	1.83 (1.68 - 2.13)
5	Kolbach Index	39.1 (33.0 - 43.0)	34.6 (31.0 - 40.0)	38.3 (32 - 46)	32.0 (29 - 34)
6	Viscosity (mPas)	1.61 (1.48 - 1.87)	1.73 (1.57 - 2.16)	1.67 (1.54 - 1.94)	1.91 (1.73 - 2.14)
7	Friability (%)	84 (66 - 93)	76 (58 - 87)	81 (68 - 92)	46 (35 - 58)
8	Malt Modification (%)	90 (73 - 96)	86 (71 - 93)	84 (73 - 94)	65 (51 - 76)
9	Homogeneity (%)	74 (63 - 85)	68 (58 - 81)	71 (62 - 83)	59 (53 - 68)
10	Diastatic Power (WK)	257.5 (146 - 390)	260.1 (166 - 379)	254.7 (151 - 384)	292.5 (166 - 437)
11	Wort colour (EBC Units)	2.7 (2.2 - 2.8)	2.3 (1.9 - 3.0)	2.8 (2.5 - 3.1)	2.4 (2.2 - 2.5)
12	Nitrogen in Wort	64.5 (56 - 70)	61.5 (55 - 70)	62.5 (50 - 71)	65.4 (58 - 78)
13	ß-glucan in wort	267 (120 - 830)	389 (150 - 1200)	339 (160 - 700)	853 (650 - 1260)

Only minor differences are seen between the spring varieties in the two growing location. Considering ß-glucan in barley, it is seen that the content is higher in the spring barley varieties (both locations) than in the winter barley varieties. In malt and wort, however, the ß-glucan values are higher in the winter barley cultivars than in the spring barley cultivars, most clearly demonstrated by comparing the samples from Jutland. This suggests a higher level of ß-glucan degradation enzymes (mainly ß-glucan solubilase and ß-glucanase) in the spring barleys which, however, is considerably environmentally dependent. This agrees with the fact that growing of winter barley for malting in Denmark has been limited due to low malt modification[9].

A PCA was initially performed on the malting quality data including 13 quality parameters of all the 50 samples (30 spring barley samples and 20 winter barley samples: in all a (50x13 matrix). Three principal components were obtained, explaining 79 % of the total variation. The score plot (landscape of samples) of the first and second principal components (PC1 and PC2) are shown in Figure 1a. The plot reveals an almost clear separation of the winter barley samples grown in Jutland (□) in the right part of the plot, while the winter barley varieties grown in Zealand (■) and all spring barley varieties (both ● and ○) appear almost in one group. From the plot it could be further concluded that the samples from Jutland (○ and □) are spread more than the samples grown in Zealand (● and ■), indicating that the growing conditions in Jutland induce larger variation in malting quality. This is especially pronounced for the winter barley samples.

The corresponding loadings plot for PC1 and PC2 is depicted in Figure 1b. The first component (x-axis) is from the left to the right mainly expanded by a cluster (a) of parameters, namely Kolbach index (KOLBACH), modification (MODIF), malt homogeneity (HOM), friability (FRIAB) and extract (EXTRACT) balanced by a cluster (b) of ß-glucan in malt (BGIM), ß-glucan in wort (BGIW) and viscosity (VISCOS) to the right. Variables within the (a) or (b) clusters correlate positively, while the correlations between a variable from cluster (a) and a variable from cluster (b) are negative. In other words, the first principal component, representing 54 % of the variation in the data Figure 1b, can be described as a functional factor mainly related to malt modification and its further consequences, expanded by the two highly intercorrelated groups. Samples to the left in Figure 1a display good modification behaviour with a high degree of modification, high friability and high proteolytic activity in terms of Kolbach Index (soluble to total nitrogen ratio), while samples to the right samples displaying inferior modification behaviour in terms of high levels of ß-glucan in malt and wort and the consequently high viscosity level. Thus, PC1 could be identified and named as the malt modification component, however, also including contributions from other parameters such as extract and Kolbach Index.

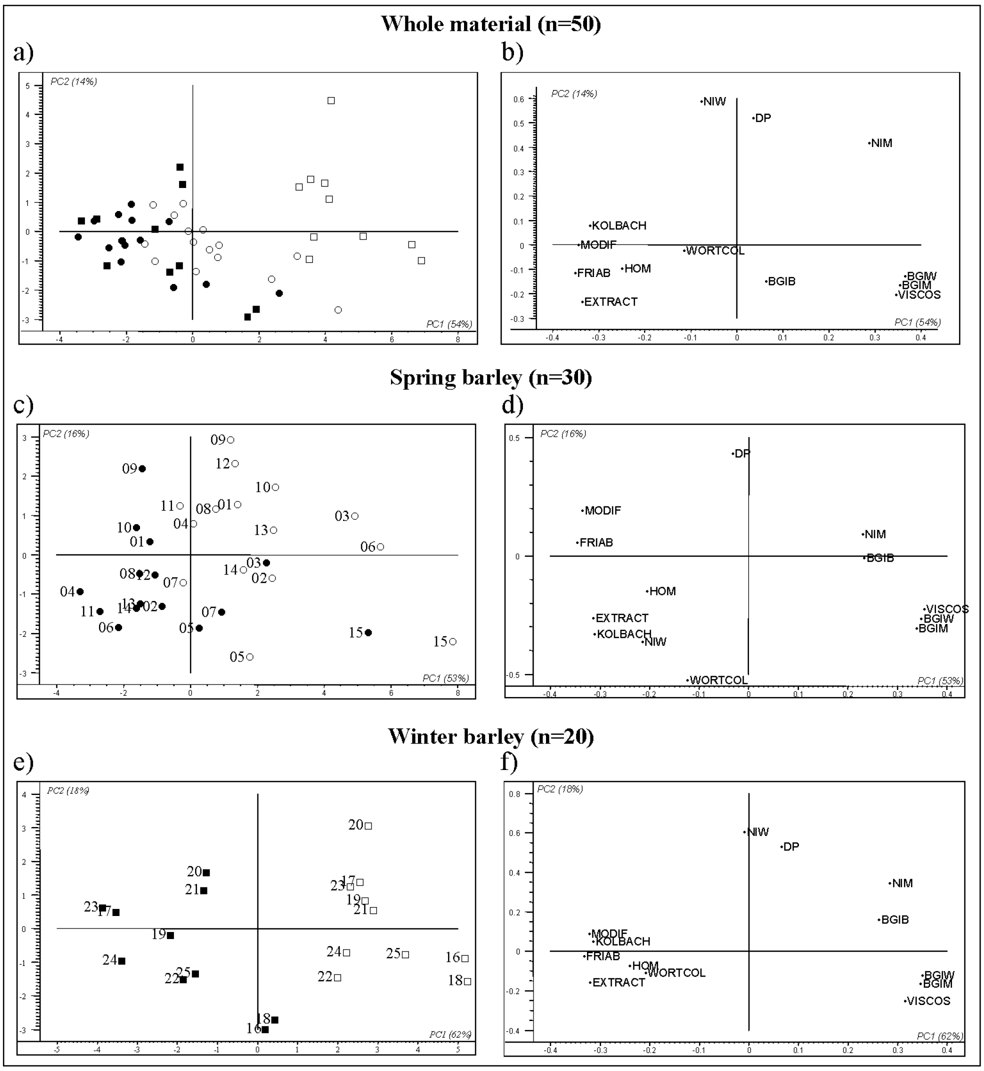


Figure 1. Score and loading plots of PCA on the 13 malting barley quality parameters for all 50 samples (Figure 1a and 1b), spring barley (Figure 1c and 1d) and winter barley (Figure 1e and 3f). The spring barley samples are marked ● (Zealand) and ○ (Jutland) and the winter barley samples are marked ■ (Zealand) and □ (Jutland). For variable names, see text.

In Figure 1b it is seen that the second component PC2 is mainly described by enzymatically influenced parameters such as nitrogen in malt (NIM), nitrogen in wort (NIW) and diastatic power (DP) in the upper part of the plot, with the two latter in the same part of the plot indicating covariance. Since principal components are by definition orthogonal, the PC2 is highly independent of the malt modification component (PC1).

The extract (EXTRACT) loading is situated in the lower left part of the plot and the loading for nitrogen in malt (NIM) in the upper right part of the plot. These two parameters seem, therefore, to contribute as a combination of the first and second principal component, and are, as expected, clearly negatively correlated. The third component (not shown in this two-dimensional plot) is in this material represented by ß-glucan in barley and wort colour, and is mainly independent of the two first components when considering the whole material of spring and winter barleys.

When combining the scores (Figure 1a) with the loadings (Figure 1b) it is seen that the winter barley samples grown in Jutland (□) are the samples with a relatively inferior malting behaviour represented in the right part of the plot. Thus, the spring barley (both locations) and the winter barley grown in Zealand represent good malting behaviour as situated in the lower left part of the plot.

In order to explore lesser trends and differences, the spring and winter barley samples were analysed in separate PCA's. A PCA was performed on the 30 spring barley samples, and the two first components are shown as scores (Figure 1c). The first three principal components comprise 77% of the total variation. By considering the loadings plot first (Figure 1d) it is seen that PC1 is nearly the same as for the whole material, but now the PC2 has changed. In contrast to the whole material (Figure 1b), the second PC is now independent of nitrogen in malt (NIM) and mainly described by diastatic power (DP) (upper part), however with a slight inversely correlated contribution from wort colour (WORTCOL) in the lower part of the plot. Nitrogen in malt and, to some extent, nitrogen in wort are now in PC3 (not shown).

The score plot corresponding to the spring barley samples in Figure 1c includes the numbers of the varieties according to Table I. In general, the Zealand samples (●) are situated in the lower left part of the plot compared to the Jutland samples (○), although the two groups are not entirely separated. The Jutland location seems to give a higher diastatic power, but lower extract. It is further seen that sample No. 15 (variety Anni) grown at both locations has an inferior maltability, as indicated by its position in the lower right corner. Its location in the score plot fits very well with the fact that this variety was never used in Denmark due to a very poor malt modification[9].

The score plot (Figure 1c) might also be used to evaluate the environmental effect on malting quality for a given genotype by considering the differences in the plot position of a given variety from one location to another. Such changes in position in the score plot vary both in *distances* and in *directions*. For example, some varieties display a large distance in the PCA plot going from one location to another, as seen when comparing sample ●06 (Miralix grown in Zealand) with ○06 (Miralix in Jutland). Other varieties seem more independent of the growing location, as seen for samples 01 (Alexis) and 05 (Caminant), indicated by smaller movements in the score plot for these samples. For most of the spring barley varieties the environmental difference is expressed in an upper right movement in the plot, such as for variety 10 (Delibes) and variety 12 (Mentor), with the latter in a more vertical direction than variety 10. Only a few varieties show a change in position to the lower right, such as variety 05 (Caminant). Sample 07 (variety Texana) seems to have a unique response to the two growing locations, as the malting quality is slightly superior (●07 more to the right) when grown in Jutland, as compared to Zealand (○07 more to the left).

The PCA on the 20 winter barley samples is shown as scores (Figure 1e) and loadings (Figure 1f). The first three principal components comprise 87% of the total variation. The loadings plot (Figure 1f) reveals a pattern close to that of the whole material (Figure 1b), although with changes in the loadings for ß-glucan in barley and wort colour. The corresponding score plot Figure 1e shows a clear-cut differentiation between the two locations. The environmental differences of the winter barley genotypes seem quite constant and pronounced. All the varieties have shifted substantially to the right and, except for variety 21, the shift is in a slightly upward direction. The general effect of a shift in location from Zealand to Jutland is similar to the spring barleys; however, for the winter barley it is much more pronounced.

We have now demonstrated the use of multivariate data analysis, in terms of PCA, for a detailed graphically based discussion with regard to varieties, growing locations, the malting quality parameters and their interaction. In an investigation involving 186 commercial malts of four spring barley varieties Munck[3] found three principal components representing 1) chemistry (enzyme activity and starch/extract), 2) physics (malting resistance) and 3) nitrogen in malt. In the current investigation the data structure and the covariance pattern and thereby the principal components are somewhat differently combined; however, the functional factor of endosperm modification and its consequences plays a major role as a functional parameter in both investigations. The difference may be explained by the fact that the present investigation includes several more genotypes representing both winter and spring barley types.

In this analysis we have seen that the first principal component (malt modification component) reveals almost the same loading pattern for the whole material, the spring barleys and the winter barleys. In contrast, the consecutive components (PC2 and PC3) involving ß-glucan in barley, nitrogen in malt, diastatic power, wort colour and nitrogen in wort seem to vary considerably between the sub-materials. The clearest difference is seen in the correlation between diastatic power and nitrogen in malt and wort, as seen by comparing PC2 in the figures 1d and 1f. For the spring barleys the correlation between diastatic power and nitrogen in malt is r=-0.06 and it is r=-0.14 for diastatic power and nitrogen in wort. For the winter barleys these two correlations are r=0.49 and r=0.68, respectively. Diastatic power is a measure of potential of amylotic power mainly contributed by α- and ß-amylase. Inactive ß-amylase is found in the ungerminated barley, while α-amylase is synthesized during malting. A positive correlation between ß-amylase, diastatic power and nitrogen in malt has been previously reported by Yan *et al.*[10]. However, our results suggest that the correlations between diastatic power, nitrogen in malt and wort are different in the investigated spring versus winter barleys.

The fact that diastatic power in this material is only partly correlated to extract is shown, since the extract is mainly described in first PC1 and diastatic power mainly described in PC2. This indicates that the extract yield is more dependent on physical parameters restricting the migration of the enzyme and accessibility to the starch than on the amount of amylotic power. This is in agreement with the findings of Brennan *et al.*[11].

Prediction of wort quality by PLSR from barley and malt quality data:

We have shown that PCA can reduce the complexity of the malt quality data for the purpose of exploratory interpretation. The loading plots in Figure 1 show that some of the measured parameters are highly intercorrelated and some therefore might be redundant. This could be utilised in order to reduce the number of malting barley quality analyses in plant breeding selection for economy and efficiency purposes. It would be interesting to utilise the barley and malt data for the prediction of wort quality in order to omit the time-consuming mashing step. This is done by applying PLSR, in which the parameters in **X** are easily attainable data such as barley and malt quality data, and where **y** is quality parameters that are more expensive and time-consuming to attain, for example, wort parameters. Figure 2a shows a predicted versus measured plot of a PLSR model using ß-glucan in barley, ß-glucan in malt, nitrogen in malt, friability, homogeneity, modification and diastatic power as **X** for the prediction of extract (**y**) as analysed in the final wort. As can be seen, a reasonable model using two components is obtained, having a correlation coefficient (r) of 0.91 and a cross-validated prediction error (RMSECV) of 0.62 corresponding to a relative error of 9.1%. In order to evaluate the importance of the different **X** variables, i.e. which of the barley and malt parameters in **X** that is most important in predicting a given wort quality parameter, the regression coefficient of the PLSR model for extract (Figure 2a) is considered in Figure 2b. The significant variables (shown in black) are estimated using the modified Jack-knife technique[8], as referred to in Material and Methods. In this investigation we use "leave-one-out" validation and for each of the 51 sub-models (one for each sample left out and one with all samples included) we get estimates of each regressor in the regression coefficient by which the mean (magnitude) and the uncertainty can be calculated and used to find the significant variables in **X**.

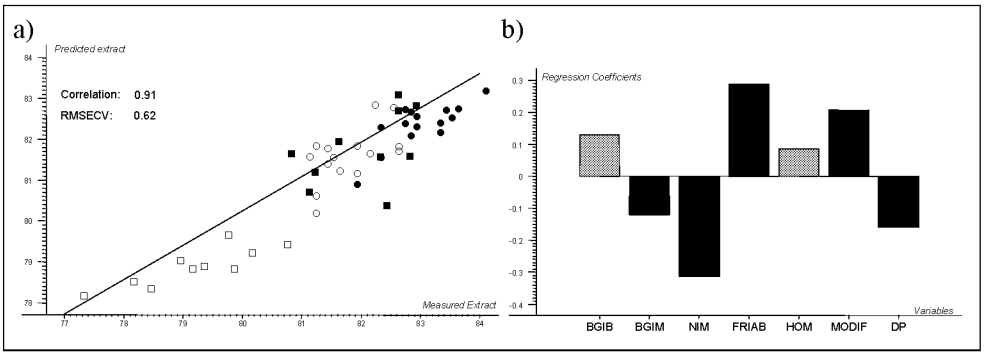


Figure 2. a) Predicted versus measured plot and regression line of a PLSR model using the combined barley and malt data as **X** for the prediction of extract yield. b) Regression coefficient with significance determination using Jack-knife validation (significant variables in black and non-significant in grey)

The PLSR regression coefficients for the extract model are shown in Figure 2b in which the significant regressors (variables) are shown in black and the non-significant variables are shown in grey. As can be seen, nitrogen in malt (negative) and malt friability (positive) are the most important malt quality parameters in predicting wort extract yield. The first parameter is mainly related to reduced starch content, while the latter is dependent on hardness – a resistance parameter in malting. Moreover, ß-glucan in malt, malt modification and diastatic power are also significant contributors to the model, although at a lower level, while ß-glucan in barley and malt homogeneity are not significant in this model.

Table III. PLSR models using barley and malt quality (X) for the prediction of wort quality data (y).

Predicted Variable (**y**)	**X** variables[a]	# PLSR	Correlation coefficient	RMSECV	RE (%)
EXTRACT	**NIM**, **FRIAB**, **MODIF**, **DP**, BGIB, **BGIM**, HOM	2	0.91	0.62	9.1
KOLBACH	**FRIAB**, **MODIF**, **BGIM**, **NIM**, **HOM**, BGIB, DP	1	0.77	2.58	15.2
VISCOS	**BGIM**, **MODIF**, **FRIAB**, **DP**, BGIB, HOM, NIM	3	0.91	0.07	10.3
WORTCOL	-	-	-	-	-
NIW	**NIM**, **MODIF**, **BGIM**, HOM, BGIB, FRIAB, DP	4	0.62	4.42	15.8
BGIW	**BGIM**, **MODIF**, **FRIAB**, **BGIB**, HOM, DP, NIM	3	0.98	65	5.7

[a] Variables listed according to importance (magnitude); significant variables in bold
- Model information excluded due to low correlation coefficient (r<0.6)

Table III summarises all the PLSR models in which barley and malt data are used for wort quality predictions. The **X** variables are listed (in order) according to the magnitude of their influence and the significant variables are marked in bold. Good predictions are achieved for extract, viscosity (VISCOS) and ß-glucan in wort (BGIW), while the barley and malt data could not predict the wort colour (WORTCOL) and only a weak prediction was achieved for Kolbach (KOLBACH) and nitrogen in wort (NIW). Thus, for prediction of extract yield, viscosity and ß-glucan in wort for screening purposes in a plant breeding material, these results suggest that the mashing step might be omitted.

Conclusions

Complex malting quality profiles of spring and winter barley consisting of 13 quality parameters have been combined and reduced into a few latent factors using Principal Component Analysis (PCA). The malt modification plays a major role in both spring and winter barleys. The spring barley and winter barley samples display different covariate latent structures, mainly reflecting differences in the nitrogen and diastatic power patterns. We have furthermore shown that graphic display as facilitated by exploratory data analysis can be utilized in order to evaluate genotype-environmental interactions by considering the position and movements of the individual objects (here genotypes) in the score plots. Thus, the samples can be individually evaluated and the corresponding loadings can be used to validate the genetic and environmental effect of a given sample in a quality perspective.

Several of the investigated malting quality parameters seem to be highly intercorrelated. This fact is utilized by applying Partial Least Squares Regression (PLSR) on barley and malt data for the prediction of wort quality in order to exclude the mashing step. This approach was successful for the modification-dependent wort parameters: extract, ß-glucan in wort and viscosity. The parameters related to protein degradation (Kolbach Index and nitrogen in wort) and wort colour were, however, more difficult to describe by barley and malt predictors. Even though these PLSR models were inferior and might not be usable directly in the laboratory, the Jack-knife validated regression coefficients might still be used for interpretation purposes, i.e. evaluate the influence of the different barley and malt parameters on the final wort of a given data set. It would be worthwhile in screening for quality in malting barley to build a broader and more diverse database than that represented in our limited example, in order to obtain more reliable models for prediction of wort quality from barley and/or malt analyses.

Acknowledgements

The authors are grateful to Carlsberg Research Laboratory for providing the barley and micro-malted samples and for performing the malt quality analyses. We also wish to thank Dr. Jørgen Larsen, Carlsberg Research Laboratory and Associate Professor Lars Nørgaard, Food Technology for constructive discussions regarding the manuscript. The authors thank The Directorate for Food, Fisheries and Agri Business for financial support through the Danish Cereal Network-B4.

References

1. Reiner, L. Die Faktorenanalyse in der Brautechnologischen Forschung. Brauwissenschaft 24 (1971) 403-412.

2. Jacobsen, T. New aspects of chemical analysis: Chemometrics - variations in trace elements in malt as shown by factor analysis. Brygmesteren 6 (1980) 149-158.

3. Munck, L. Quality criteria in the production chain from malting barley to beer. Ferment 4 (1991) 235-241.

4. Nielsen, J. P. Evaluation of malting barley quality using exploratory data analysis. II. The use of kernel hardness and image analysis as screening methods. Journal of Cereal Science, submitted (2002).

5. Analysis by the European Brewery Convention. Brauerei- und Getränke-Rundschau, Zurich (1987).

6. Wold, S., K. Esbensen, and P. Geladi. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 2 (1987) 37-52.

7. Martens, H and Næs, T. 1993. Multivariate Calibration. Wiley, New York.

8. Martens, H. and Martens, M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Quality and Preference 11 (2000) 5-16.

9. Larsen, J. 2002. Personal communication. Carlsberg Research Laboratory.

10. Yan, X., J. Zhu, S. Xu, and Xy, Y. Genetic effects of embryo and endosperm for four malting quality traits of barley. Euphytica 106 (1999) 27-34.

11. Brennan, C. S., Amor, M. A., Harris, N., Smith, D., Cantrell, I., Giggs, D. and Shewry, P. R. Cultivar Difference in Modification Patterns of Protein and Carbohydrated Reserves during Malting of Barley. Journal of Cereal Science 26 (1997) 83-93.

Paper 2

Evaluation of malting barley quality using exploratory data analysis. II. The use of kernel hardness and image analysis as screening methods

Jesper Pram Nielsen

Abstract

This paper presents an exploratory investigation of the use of image analysis and hardness analysis of barley kernels for characterisation and prediction of malting quality. A sample set of fifty barley samples representing 15 spring barley and 10 winter barley varieties grown at two locations in Denmark was used. The samples were micro-malted and mashed and analysed for 13 quality parameters according to the official methods of the European Brewery Convention. A sub-sample of the barley samples was analysed on two different single kernel instruments: 1) Foss Tecator GrainCheck was applied for non-destructive recording of single kernel size and shape (width, length, roundness, area, volume and total light reflectance) and 2) Perten Single Kernel Characterization System 4100 was applied for single kernel hardness and weight determinations. The eight variables from these single seed analyses have been used in two different ways, either as means and standard deviations, or as appended histogram spectra representing 250 kernels from each bulk sample. By the two methods, it has been possible to obtain reasonable Partial Least Squares Regression (PLSR) models for the structural and physical part of the malting quality complex associated to malt modification, but it was as expected impossible to predict the biochemical parameters associated with nitrogen chemistry and enzymatic power. The best model was achieved for ß-glucan in barley. The hardness of the barley kernels is by far the most important variable for describing malting performance. The additional use of the morphological data as acquired by fast non-destructive image analysis, however, also reveals some malting quality information by improving the calibration models based on hardness alone. The brightness of the kernels is by far the most important GrainCheck variable but also kernel size and shape is associated to malting performance. In general, the utilisation of the single kernel readings (used as histogram spectra), compared to sample mean and standard deviation, did not provide additional information for an improved prediction of the malting quality parameters.

Introduction

Testing of barley for malting quality in barley breeding programmes is an expensive and time-consuming task involving micro-malting and mashing and followed by several slow chemical analyses. Predicting the malting quality directly from the barley samples by fast screening methods is therefore of interest to plant breeders for early quality screening and selection. The attempt to predict malting quality either as measured directly on the malt or without actually malting the barley sample has been extensively studied for many years, mainly by means of near infrared spectroscopic analyses.

Kernel size and shape contain information relevant for the end-use quality of cereals in general. Thus, image analysis has become a promising analysis for the cereal industry. Digital images of kernels can be used to estimate kernel characteristics such as size, shape and colour. This is a fast and objective analysis of grain morphology, which can be considered as an advanced version of the first "quality control" during harvest, when the farmer picks up a handful of kernels in the combine and visually evaluates size, shape and brightness of the grains.

Several commercial automated instruments based on this technique are now available for example for purity analyses. In wheat, the technique has been used to discriminate classes[1] and to screen for milling yield[2]. Gebhardt *et al.*[3] employed digital image analysis for quantifying kernel morphology variation in six-row barley in relation to malting quality, Ninomiya *et al.*[4] have used the method for evaluation of wrinkles on husks of malting barley, and García del Moral *et al.*[5] have recently employed image analysis on barley kernels as a predictor for malting quality.

For wheat, it is further found that kernel hardness is an important characteristic, as an indicator of the ease of processing in the milling process, and of the end-use quality, for example, with respect to starch damage and flour yield. It is also known that barley endosperm texture affects the malt modification process during malting by affecting water uptake and consequently enzyme synthesis and movement within the endosperm[6]. Recently, Andersson *et al.*[7] studied the variation and correlation between chemical and physical characteristics of barley samples including kernel hardness, but found only a low correlation between kernel hardness and physical and chemical grain properties.

The purpose of this investigation is to use multivariate data regarding barley seed size, shape and hardness as a screening method for malting barley quality. This is done by performing image analysis and kernel hardness analysis (SKCS) on spring and winter barley samples grown in Denmark. The same sample set was used by Nielsen and Munck (2002) in an exploratory study defining the functional factors of classical malting barley data[8]. In this investigation the samples are analysed in bulk, but the applied instruments conduct single kernel readings, and thus provide data on each of the single kernels in the sample. Since the kernel-to-kernel variability in the barley used for malting is of great importance for the malt quality, the single kernel output of the applied instruments is utilised in order to evaluate their use in malting barley characterisation.

Experimental

Sample collection:

The barley samples originate from trials under the European Brewery Convention (EBC) harvested in 1995. Fifteen spring barley varieties and ten winter barley varieties were grown at two different locations in Denmark, Jutland and Zealand, providing 50 malting barley samples in total[8]. The barley grain samples were screened over a standard 2.5-mm sieve and the grains above 2.5 mm were subjected to the micro malting procedure.

Malting quality analyses:

The following 13 quality analyses were performed according to the official methods of the European Brewery Convention[9]: ß-glucan in barley (BGIB) and in malt (BGIM), extract, nitrogen in malt (NIM), Kolbach Index (KOLBACH), viscosity (VISCOS), friability (FRIAB); malt modification (MODIF), malt homogeneity (HOM), diastatic power (DP), wort colour (WORTCOL), nitrogen (NIW) and ß-glucan in wort (BGIW) (see Table I variable number 1 to 13).

Barley grain morphology measurements:

Barley grain morphology was measured by digital image analysis using a GrainCheck^{TM} 310 instrument (FossTecator, Höganas, Sweden). More than 250 barley kernels of the screened barley grain samples were fed onto a conveyer belt, from which the kernels were automatically imaged by a RGB video camera. The multi-kernel images were segmented into single kernel images from which several morphological variables were automatically assessed.

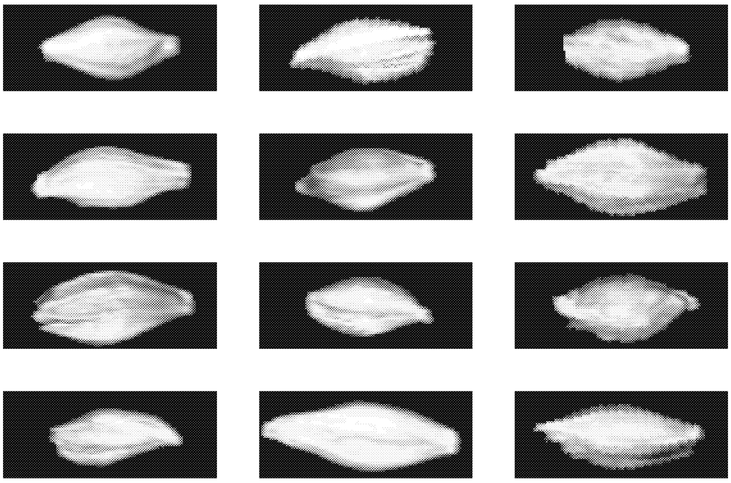


Figure 1. Examples of twelve single-kernel GrainCheck images, from which the registered morphological data was estimated. The depicted kernels show considerable variations in size, shape and reflected light intensity.

Figure 1 shows an example of GrainCheck images of single kernels with differences in size, shape and brightness. The instrument is specifically designed for purity analysis of grain samples based on morphology and colour data. In this investigation the following variables were exported from the instrument: kernel width (WIDTH), kernel length (LENGTH), roundness (ROUND), area (AREA), volume (VOLUME), and total light reflectance (INT) (see Table I variable numbers 14 to 25). These variables were exported from the first 250 single kernels of an analysed sub-sample of each of the 50 bulk samples.

Hardness analysis:

A Single Kernel Characterization System (SKCS) 4100 (Perten Instruments Inc., Reno, NV, USA) was used for assessment of single barley kernel hardness and kernel weight. The SKCS is designed for hardness analysis in wheat, and the Hardness Index (HI) reported from the instrument is based on the algorithm used to

Table I. Mean and ranges (in parentheses) of the samples of the quality, morphology and hardness data

	ID	Spring barley Zealand (●) n=15	Spring barley Jutland (○) n=15	Winter barley Zealand (■) n=10	Winter barley Jutland (□) n=10
Malting quality data:					
1	ß-glucan in barley (%)	3.87 (3.14 - 4.60)	4.05 (3.49 - 4.98)	3.09 (2.77 - 3.42)	3.60 (3.20 - 3.92)
2	ß-glucan in malt (%)	0.48 (0.21 - 1.52)	0.65 (0.26 - 1.79)	0.50 (0.21 - 1.10)	1.24 (1.00 - 1.80)
3	Extract yield (%)	82.8 (81.7 - 83.9)	81.6 (80.9 - 82.4)	81.8 (80.6 - 82.7)	78.9 (77.1 - 80.5)
4	Nitrogen in malt (%)	1.47 (1.35 - 1.55)	1.60 (1.49 - 1.71)	1.46 (1.33 - 1.74)	1.83 (1.68 - 2.13)
5	Kolbach Index	39.1 (33.0 - 43.0)	34.6 (31.0 - 40.0)	38.3 (32 - 46)	32.0 (29 - 34)
6	Viscosity (mPas)	1.61 (1.48 - 1.87)	1.73 (1.57 - 2.16)	1.67 (1.54 - 1.94)	1.91 (1.73 - 2.14)
7	Friability (%)	84 (66 - 93)	76 (58 - 87)	81 (68 - 92)	46 (35 - 58)
8	Malt Modification (%)	90 (73 - 96)	86 (71 - 93)	84 (73 - 94)	65 (51 - 76)
9	Homogeneity (%)	74 (63 - 85)	68 (58 - 81)	71 (62 - 83)	59 (53 - 68)
10	Diastatic Power (WK)	257.5 (146 - 390)	260.1 (166 - 379)	254.7 (151 - 384)	292.5 (166 - 437)
11	Wort colour (EBC Units)	2.7 (2.2 - 2.8)	2.3 (1.9 - 3.0)	2.8 (2.5 - 3.1)	2.4 (2.2 - 2.5)
12	Nitrogen in Wort (mg/100 ml)	64.5 (56 - 70)	61.5 (55 - 70)	62.5 (50 - 71)	65.4 (58 - 78)
13	ß-glucan in wort (mg/l)	267 (120 - 830)	389 (150 - 1200)	339 (160 - 700)	853 (650 - 1260)
Barley kernel morphology data:					
14	Kernel width (mm)	3.7 (3.6 - 3.8)	3.7 (3.6 - 3.9)	3.5 (3.4 - 3.6)	3.6 (3.5 - 3.7)
15	Width homogeneity[b]	0.21	0.24	0.22	0.23
16	Kernel length (mm)	8.2 (7.9 - 8.8)	8.5 (8.0 - 9.2)	8.4 (8.1 - 8.8)	8.9 (8.4 - 9.3)
17	Length homogeneity[b]	0.65	0.77	0.64	0.86
18	Kernel roundness (AU)[a]	0.3 (0.30 - 0.35)	0.31 (0.27 - 0.34)	0.30 (0.29 - 0.31)	0.28 (0.27 - 0.31)
19	Roundness homogeneity[b]	0.05	0.06	0.04	0.06
20	Kernel area (mm^2)	21.8 (20.9 - 23.4)	22.5 (20.9 - 24.1)	21.6 (20.3 - 23.0)	23.1 (21.5 - 24.7)
21	Area homogeneity[b]	2.45	2.86	2.30	2.79
22	Kernel volume (mm^3)	50.9 (47.9 - 54.9)	52.2 (47.7 - 57.1)	48.8 (44.4 - 52.7)	53.0 (47.7 - 57.9)
23	Volume homogeneity[b]	7.98	9.56	7.50	8.68
24	Total light reflectance (Int)	74.5 (71.9 - 76.1)	76.3 (74.3 - 77.8)	70.8 (68.1 - 73.5)	71.2 (70.0 - 73.0)
25	Int. homogeneity[b]	4.55	3.84	4.11	5.32
Barley kernel texture and weight:					
26	Relative Hardness Index (RHI)	56.0 (48.3 - 77.2)	64.1 (56.0 - 84.2)	49.7 (43.1 - 59.0)	67.3 (57.4 - 76.2)
27	RHI homogeneity[b]	12.8	11.4	13.4	11.6
28	Average Kernel weight (mg)	50.2 (47.1 - 54.6)	50.0 (46.8 - 54.2)	48.3 (42.6 - 53.5)	51.9 (47.3 - 56.6)
29	Weight homogeneity[b]	7.2	8.1	6.6	7.7

[a] Values in the range of 0 – 1. A perfect circle has roundness 1, while a very narrow elongated object has roundness close to 0.

[b] Sample homogeneity calculated as standard deviation between the 250 single kernels analysed

define differences in wheat hardness and therefore may not be appropriate for barley. The HI value is based on the amount of force required to crush the kernels (area integration of the crush profile) corrected by kernel weight and moisture. Since no apparent hardness standard or reference method has been established for barley, the hardness will be reported here as a relative hardness index (RHI). The Relative Hardness Index (RHI) and kernel weight (WEIGHT) (see Table I variable numbers 26 to 29) variables were exported from each of the first 250 single kernels of an analysed sub-sample of each of the 50 bulk samples.

Exploratory data analysis:

The methods employed are described in detail in the literature, and a brief description was given by Nielsen and Munck (2002)[8]. Principal Component Analysis (PCA)[10] and Partial Least Squares Regressions (PLSR)[11] were performed using Unscrambler version 7.6 SR-1 (CAMO A/S, Norway). The PCA results are shown as score, loading- and bi- (combining scores and loadings) plots of the analysed **X** matrix containing the data from the two instruments. The PLSR models are presented as correlation coefficients (r), root mean square error of cross-validation (RMSECV) and relative prediction errors (RE) calculated as RMSECV/range. The importance of the **X** variables is evaluated using Jack-knife validation proposed by Martens and Martens (2000)[12].

Results and Discussion

Kernel morphology and hardness for malting barley characterisation:

In the following, data based on individual barley kernels will be used in two different ways, namely as sample histograms or as means and standard deviations. It should be emphasised that the present investigation on a limited material will allow for an investigation of the potential usefulness of the methods and not for development of global calibrations.

Single kernel data as appended histograms:

The chosen methods for analysis of morphology and hardness are based on readings of the individual kernels in each sample, and even if these methods are usually used to give average values on a bulk sample only, the advantage of having single kernel data will be utilised by performing PCA on "appended histograms" of the individual single kernel readings. For a given parameter, the minimum and maximum single kernel reading is found in the data matrix of all 250 single kernels readings of the 50 samples, i.e. 12500 single kernel readings, and the found range is divided into 15 equidistant intervals. For each sample and parameter, a histogram is made on the basis of 250 single kernel readings. Thus, for each sample, eight histograms are made, one for each recorded parameter, thereby giving 8x15 equal to 120 variables for each sample. The appended histograms (shown as curves) for the 50 samples are given in Figure 2a. A PCA score plot (Figure 2b) of these mean centred histogram spectra only reveals unclear grouping according to barley type [winter (□ and ■) versus spring (○ and ●)] or growing location [Jutland (white symbols) and Zealand (black symbols)].

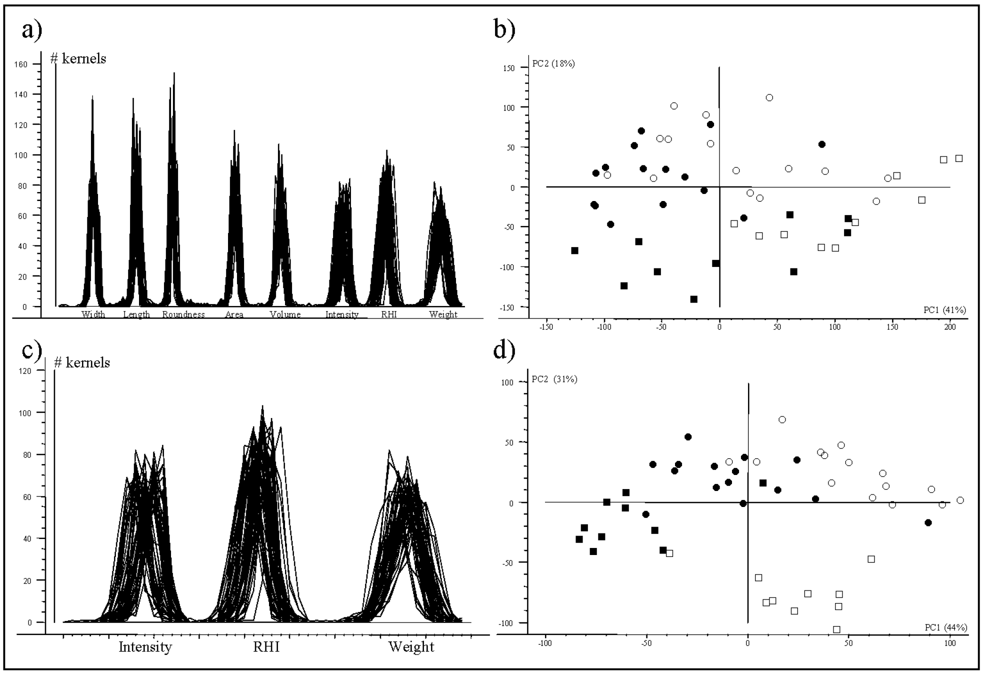


Figure 2. (a) Appended histogram spectra of all recorded characteristics (a). The histograms from left to right are: width, length, roundness, area, volume, intensity, hardness (RHI) and weight. (b) PCA score plot of the spectra in a). The spring barley samples are marked ● (Zealand) and ○ (Jutland) and the winter barley samples are marked ■ (Zealand) and □ (Jutland). (c) Enlarged appended histogram spectra of intensity, RHI and weight. (d) PCA score plot of the spectra in c).

As seen from the histogram spectra (Figure 2a), the largest variations are in the histograms of light reflected intensity (INT) from the kernel, in the Relative

Hardness Index (RHI) and the kernel weight. These parts of the histogram spectra are enlarged in Figure 2c and a PCA on them is shown in the score plot in Figure 2d. Except for a few samples, an almost clear differentiation between the four groups is seen. In particular, the winter barleys grown in Jutland (□) stand out as a separate group. This grouping was also demonstrated in a PCA on the malting quality data of the same samples[8]. We could therefore preliminarily conclude that the physical readings in terms of histograms should be considered for a malting barley quality characterisation.

The above analysed histograms contain information on the different kernel characteristics regarding both *average* values and the *homogeneity* within a sample. In the further analysis the importance of the different kernel parameters will be elaborated in order to differentiate between the average and the homogeneity values that are responsible for differentiation of the analysed malting barley samples.

Single kernel data as mean and standard deviations:

The morphology and hardness data are listed in Table I (variables 14-29) presented as sample mean and sample homogeneity (standard deviations of the 250 single kernels) for the four groups, namely the winter and spring barley grown in Jutland and Zealand, respectively. A PCA was computed on the mean and standard deviations separately, i.e. 8 variables in each. The biplot combining the score and loading plots of this PCA is shown in Figure 3a displaying PC1 and PC2. As can

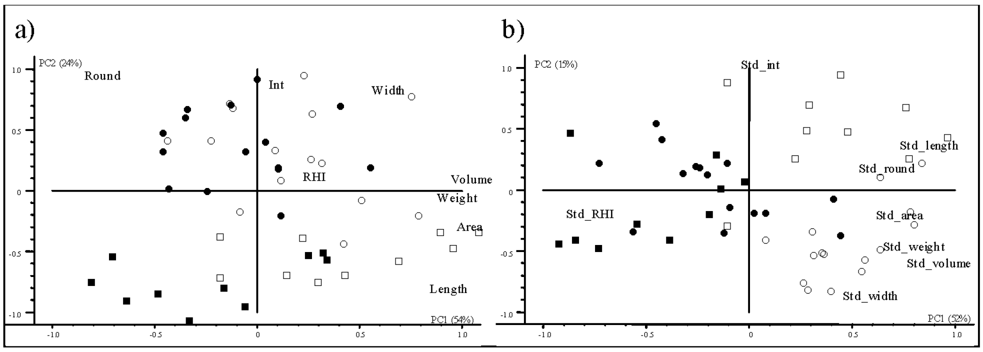


Figure 3. Bi-plots (scores and loadings) of PCA on mean values (a) and standard deviations (b) for all the recorded morphology and hardness variables. The spring barley samples are marked ● (Zealand) and ○ (Jutland) and the winter barley samples are marked ■ (Zealand) and □ (Jutland).

be seen, the morphological data are highly correlated, demonstrated by a cluster mainly representing size variables (WEIGHT, VOLUME, AREA and LENGTH) in the first PC direction (x-axis). No clear differentiation is seen between the four sample groups in this direction. The second PC is mainly expanded by the parameters roundness (ROUND), total light reflectance (Int) and kernel width (WIDTH) grouped in the upper part, and kernel length (LENGTH) in the lower part of the plot. This component clearly separates the spring barley samples (○ and ●) in the upper part of the plot from the winter barley samples (□ and ■) in the lower part. The third PC mainly represents kernel hardness (RHI), however with no significant sample groupings (plot not shown).

The corresponding PCA on the sample homogeneity data (i.e. on the standard deviations of the 250 single kernels) was computed in which four significant PC's were found, comprising 98% of the variation. A biplot of PC1 and PC2 is shown in Figure 3b. Except for a few samples a clear clustering of the Jutland (white symbols) versus Zealand (black symbols) samples is seen along the first PC. From the loadings it is seen that this component is expanded by the RHI homogeneity to the left and the homogeneity of the remaining parameters to the right. Thus, compared to the Zealand samples, the Jutland samples are more inhomogeneous on all measured parameters except for the relative hardness. The reason for the considerably higher hardness inhomogeneity in the Zealand samples and the clear differentiation between the two growing locations based on homogeneity data is unclear. The material in this investigation should, however, be expanded in order to elaborate these results.

Exploring the link between morphology, hardness and malting barley quality data:

Principal Component Analysis:

In order to investigate the link between the morphological and hardness data on one hand and the malting barley quality data on the other hand, a PCA was computed on the 50 samples using all the 29 variables listed in Table I. Figure 4 shows a loading plot of PC1 and PC2 of this PCA explaining 55 % of the total variation in the data. By first considering the malting barley parameter, it is seen that the malt modification component as discussed by Nielsen and Munck[8] is still clearly evident, however here as a combination of PC1 and PC2 going from the upper left to the lower right in the plot. The kernel roundness (ROUND) is situated near extract (EXTRACT), modification (MODIF), homogeneity (HOM), friability

(FRIAB) and Kolbach (KOLBACH) variables. Kernel roundness thus seems to be indicative of good malting behaviour (the rounder the better). To the lower right, high values of ß-glucan in malt (BGIM) and wort (BGIW), viscosity (VISCOS) and nitrogen in malt (NIM) indicate high malting resistance and low extract. Since these variables are situated in the same area as kernel length (LENGTH), hardness (RHI), length and intensity homogeneity (Std_length and Std_int), these variables seem to indicate inferior malting behaviour.

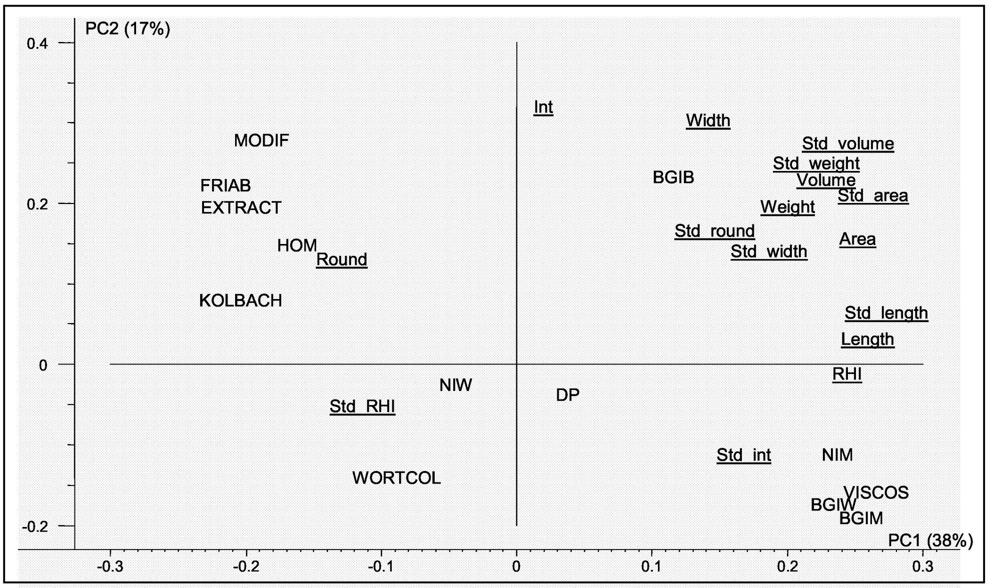


Figure 4. Loading plot of all malting barley quality, morphological and hardness variables. The malting barley quality variables (in capital letters) are: ß-glucan in barley (BGIB), malt (BGIM) and wort (BGIW), extract, nitrogen in malt (NIM) and wort (NIW), Kolbach, viscosity (VISCOS), friability (FRIAB), malt modification (MODIF), modification homogeneity (HOM), diastatic power (DP) and wort colour (WORTCOL).

From the kernel size parameters (length, area, width, volume) it is seen that the sample means and sample homogeneities (Standard deviations) are correlated, as they are situated near each other. On the other hand, this is not the case for the parameters roundness, hardness and light intensity.

Partial Least Squares Regressions:

The loading plot in Figure 4 represents a coarse map of the relationships between the morphology, hardness and malting quality data. This gives indications of which regression models based on morphology and hardness that would be expected to be useful for the prediction of malting barely quality.

Partial Least Squares Regression models were computed for predicting each of the 13 classical malting quality parameters (**y**'s) using combined morphology and hardness data as **X** variables. This was done in two different ways for each malting barley quality parameter. First, the means and standard deviations (in total 16 variables) are used as **X** for the prediction of the **y**'s. Secondly, the histogram spectra (Figure 2a) are used as **X** for prediction of the **y**'s. Table II summarises the obtained PLS prediction models for each of the 13 barley and malt quality parameters (**y**) using morphology and hardness parameters.

Table II. PLSR models based on barley kernel morphology and hardness (X) for prediction of malting quality data (y). Models marked with (MS) are based on the mean and standard deviation as X data and models marked with (HIS) are based on the histograms as X data.

Predicted Variable (**y**)	# PLS	Corr. Coeff. (r)	RMSECV	RE (%)[a]	Significant **X** variables[b]
BGIB (MS)	4	0.86	0.25	11.3	RHI, intensity, width, round, length, volume
BGIB (HIS)	3	0.84	0.26	11.8	
BGIM (MS)	4	0.74	0.30	19.0	RHI, intensity
BGIM (HIS)	5	0.75	0.30	19.0	
EXTRACT (MS)	3	0.75	1.0	14.7	RHI, intensity, std_length, round, width, length
EXTRACT (HIS)	2	0.71	1.1	16.8	
NIM (MS)	-	-	-	-	-
NIM (HIS)	-	-	-	-	-
KOLBACH (MS)	3	0.74	2,7	15.9	RHI, Std_RHI, intensity, std_width
KOLBACH (MS)	6	0.71	2.9	17.1	
VISCOS (MS)	5	0.71	0.12	17.6	RHI, intensity
VISCOS (HIS)	3	0.65	0.13	19.2	
FRIAB (MS)	4	0.82	8.94	15.4	RHI, intensity, Std_intensity
FRIAB (HIS)	3	0.81	9.27	16.0	
MODIF (MS)	5	0.84	6.1	13.6	RHI, intensity, round, length
MODIF (HIS)	3	0.78	7.01	15.6	
HOM (MS)	4	0.63	6.80	21.3	Std_RHI, RHI, round, length
HOM (HIS)	3	0.65	6.67	20.8	
DP (MS)	-	-	-	-	
DP (HIS)	-	-	-	-	
WORTCOL (MS)	-	-	-	-	
WORTCOL (HIS)	-	-	-	-	
NIW (MS)	-	-	-	-	
NIW (HIS)	-	-	-	-	
BGIW (MS)	5	0.82	170	14.9	RHI, intensity
BGIW (HIS)	5	0.80	181	15.9	

[a] Calculated as RMSECV divided by the range of the parameter in question

[b] Significant variables found by Jack-knifing, listed according to importance (magnitude).

- Model information excluded due to low correlation coefficient (r<0.6)

For a given malting quality parameter, the first model is based on mean and standard deviations (labelled MS) in **X** and the second model in based on the histogram spectra as **X** (labelled HIS). The number of PLS components (#PLS), the correlation coefficients (r), RMSECV's and relative errors (RE) are listed together with the significant **X** variables in order of importance (only for the models based on mean and standard deviation data). For the parameters nitrogen in malt (NIM), diastatic power (DP), wort colour (WORTCOL) and nitrogen in wort (NIW), it was not possible to obtain a reasonably good PLS model (correlation coefficients less than 0.6), and these models are therefore excluded from the table. For the nine remaining quality parameters, predictive models have been obtained with correlation coefficients in the range 0.63 to 0.86 and with relative errors from 11.3% to 21.3% of the range. It should be emphasised that several of these malting quality parameters are highly correlated and that some of the predictive performances might be based on these internal correlations. These preliminary results should be verified and extended by a larger calibration base with regard to number of samples as well as with regard to variation sources, i.e. more genotypes, growing location and growing years. Nevertheless, some of these performances seem to be in an acceptable range for early screening for quality in breeding programmes.

Figure 5a shows the prediction versus measured plot of one of the PLS models, namely for the prediction of ß-glucan in barley using the mean and standard deviation data as **X**. A correlation coefficient of 0.86 and a prediction error RMSECV of 0.25 is achieved, which is an encouraging result and comparable with results obtained using NIR spectroscopy[13].

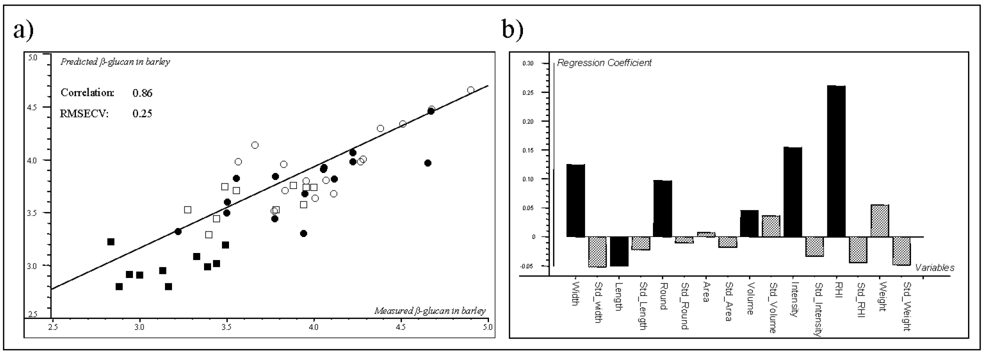


Figure 5. Predicted versus measured plot (a) and regression coefficients (b) of a PLS model based on mean and standard deviation data of barley morphology and hardness data (X) for prediction of ß-glucan in barley (y). The significant variables are shown in black the non-significant are shown in grey.

In order to illustrate the importance of the different **X** variables, i.e. which of the morphological/hardness data is most important in describing/predicting the ß-glucan in barley, the regression coefficients are considered in Figure 5b. The uncertainty of the variables is estimated using the modified Jack-knife technique as described earlier[12], and the significant variables are shown in black. As seen from Figure 5b, kernel hardness (RHI), light intensity (INT), kernel width (WIDTH) and kernel roundness (ROUND) are the most important **X** variables in describing the ß-glucan level, as shown by the highest magnitudes of the regressors. Kernel length and volume are also significant, however less important, since these show lower magnitudes. The ten remaining **X** variables depicted in grey colour are not significant for the PLS prediction of ß-glucan in barley in this material. Thus, the sample homogeneity (standard deviations) for any of the recorded variables does not contribute to describe the level of ß-glucan in barley.

With regard to the significant **X** variables (Table II) of all the PLSR models, it is seen that barley kernel hardness (RHI) plays an important role; the lower the RHI, the better malting performance. The hardness of the barley kernels is thus a good predictor for maltability, especially with respect to the physical based parameters in terms of modification, friability, and the cell wall complex including ß-glucan in barley, malt, wort and viscosity. Direct correlations between RHI and the nine malting quality parameters of which a PLSR model could be developed range from r=0.52 to r=0.71, but even though hardness as a single parameter is the most important, the range of correlation coefficients increases considerably to r values from 0.63 to 0.86 by introducing the other barley kernel morphology data, especially the light intensity, into the predictive models (Table II).

At the same time it is found that barley kernel morphology and RHI, on the contrary, are not good predictors for the quality parameters accounting for the nitrogen chemistry (nitrogen in malt and nitrogen in wort), wort colour and for the enzymatic activity (diastatic power). This is, however, not surprising. In contrast, it is interesting to note that reasonable predictive models using morphology and hardness data from barley were obtained for ß-glucan in barley, malt and wort. This could not be entirely based on indirect correlation between ß-glucan in barley on one hand and with ß-glucan in malt and wort on the other hand, since ß-glucan in barley is nearly independent of the two latter, as seen by their orthogonal directions in the loading plot in Figure 4.

The overall importance of barley kernel hardness to malting quality as shown in this investigation is in agreement with earlier investigations. Brennan *et al.*[14] found that a strong starch-protein binding is related to poor malting barleys and that good malting quality cultivars have a weak association between starch granules and protein matrix. This association was independent of the nitrogen level. Chandra *et al.*[6] recently showed that steely kernels had higher protein and ß-glucan levels and had a slower redistribution of water across the endosperm. This will slow down migration of modification enzymes during malting and result in an uneven modification and thus a lower extraction during mashing and high viscosity in the wort. The SKCS Hardness Index used in the current investigation is based on a crushing profile acquired during single seed crushing between a rotor and a crescent. The RHI of the kernels can thus be due to several physical and chemical characteristics of the barley kernel not necessarily limited to endosperm cell structure. The thickness of the endosperm cell walls might also contribute to RHI and thus can be used to predict the malting resistance parameters, but contains no information of the nitrogen chemistry.

The significance of the sample homogeneities in terms of standard deviations of the recorded characteristics only seems to play a major role for the prediction of malt homogeneity (HOM) (Table II). For this prediction, the standard deviation of the single RHI readings is the most important **X** variable, thus suggesting that automated SKCS single kernel hardness measurements might be useful for modification homogeneity in malt. However, the prediction error is far too high for practical use.

Generally, only slight differences are seen between the PLSR models based on histogram spectra versus the PLSR models based on means and standard deviations, with the latter being a little superior for most of the predictions. Thus, the utilization of the single kernel readings in terms of histograms does not improve malting quality predictions, compared to the means and standard deviations. This indicates that the within sample single kernel variation is normally distributed and that the reduction of the histograms into simple means and standard deviations prior to the data analysis seems feasible.

Conclusions

A characterisation of malting barley quality includes a range of reference parameters reflecting physical, chemical and enzymatic properties of the barley and malt. These analyses are produced by expensive, time consuming and destructive methods and the need for screening methods in barley plant breeding is therefore obvious in order to facilitate a higher sample throughput. The perfect instrument would be a non-destructive screening measurement, which by analysing the un-germinated barley samples could be able to predict all barley, malt and wort parameters. Since the conversion of barley into malt involves complex interactions between biochemical and structural parameters in the germination, predictions directly from measurements on barley could not be expected to take into account enzymatically induced parameters.

The single seed morphological data was non-destructively registered using the Foss Tecator GrainCheck and the single kernel hardness and weight was destructively assessed using the Perten SKCS 4100. Both instruments are fully automated and single kernel data of 250 kernels can be measured within a few minutes, making them of interest for screening purposes in malting barley breeding programmes. The link between barley kernel characteristics based on these two instruments and malting quality data has been investigated using exploratory data analysis. The morphological data from the single seed analysis have been used in two different ways, either as means and standard deviations or as appended histogram spectra representing 250 seeds from each bulk sample. It has been possible to obtain reasonable PLS models for the structural and physical part of the malting quality complex associated to malting resistance, but it was, not surprisingly, impossible to model the biochemical parameters associated to nitrogen chemistry and enzymatic power. The best model was achieved for ß-glucan in barley.

The SKCS Relative Hardness Index (RHI) is by far the most important variable for describing the malting performance. The SKCS instrument has a potential for early screening in barley breeding, although the method is destructive. The additional use of the morphological data as acquired by fast non-destructive image analysis also reflects some malting quality information by improving the calibration models based on RHI alone. The brightness of the kernels is by far the most important morphological variable.

The utilisation of the single kernel readings from the image and hardness instruments does not seem to provide additional information for improved prediction of the malting quality parameters, except for malt homogeneity where the RHI homogeneity was the most important variable. The prediction error of this model, however, was too high for practical use.

Further research on the use of image analysis and hardness for characterisation of malting barley quality is needed. This should include larger data sets with more extreme differences and with more samples within each variety, allowing for a more comprehensive validation of the results.

Acknowledgements

The author is grateful to Carlsberg Research Laboratory for providing the barley and micro-malted samples and for performing the malt quality analyses. He also wishes to thank Dr. Jørgen Larsen, Carlsberg Research Laboratory, Professor Lars Munck and Associate Professor Lars Nørgaard, Food Technology for constructive discussions regarding the manuscript. The author is also grateful to The Directorate for Food, Fisheries and Agri Business for financial support through the Danish Cereal Network-B4.

References

1. Zayas, I., Lai, F. S. and Pomeranz, Y. Discrimination Between Wheat Classes and Varieties by Image Analysis. Cereal Chemistry 63 (1986) 52-56.

2. Berman, M., Bason, M. L., Ellison, F., Peden, G. and C. W. Wrigley. Image Analysis of Whole Grains to Screen for Flour-Milling Yield in Wheat Breeding. Cereal Chemistry 73 (1996) 323-327.

3. Gebhardt, D. J., Rasmusson, D. C. and Fulcher, R. G. Kernel Morphology and Malting Quality Variation in Lateral and Central Kernels of Six-Row Barley. ASBC Journal 51 (1993) 145-148.

4. Ninomiya, S., Sasaki, A. and Takemura, K. Evaluation of fineness of wrinkles on husks of malting barely (*Hordeum vulgare* L.) by texture analysis of digital image data. Euphytica 64 (1992) 113-121.

5. García del Moral, L.F., Sopena, A., Montoya, J.L., Polo, P., Voltas, J., Codesal, P., Ramos, J.M. and Molina-Cano, J.L. Image Analysis of Grain and Chemical Composition of the Barley Plant as Predictors of malting Quality in Mediterranean Environments. Cereal Chemistry 75 (1998) 755-761.

6. Chandra, G. S., Proudlove, M. O. and Baxter, E. D. The structure of barley endosperm - An important determinant of malt modification. Journal of the Science of Food and Agriculture 79 (1999) 37-46.

7. Andersson, A. A. M., Elfverson, C. , Andersson, R. , Regnér, S. and P. Åman. Chemical and physical characteristics of different barley samples. Journal of the Science of Food and Agriculture 79 (1999) 979-986.

8. Nielsen, J. P. and Munck, L. Evaluation of malting barley quality using exploratory data analysis. I. Extraction of information from micro-malting data of spring and winter barley data. Journal of Cereal Science, submitted (2002).

9. Analysis by the European Brewery Convention. Brauerei- und Getränke-Rundschau, Zurich (1987).

10. Wold, S., Esbensen, K. and Geladi, P. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 2 (1987) 37-52.

11. Martens, H and Næs, T. 1993. Multivariate Calibration. Wiley, New York.

12. Martens, H. and Martens, M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Quality and Preference 11 (2000) 5-16.

13. Czuchajowska, Z., Szczodrak, J. and Pomeranz, Y. Characterization and Estimation of Barley Polysaccharides by Near-Infrared Spectroscopy. I. Barleys, Starches, and ß-D-Glucans. Cereal Chemistry 69 (1992) 413-418.

14. Brennan, C. S., Harris, N., Smith, D. and Shewry, P. R. Structural Differences in the Mature Endosperms of Good and Poor Malting Barley Cultivars. *Journal of Cereal Science* **24** (1996) 171-177.

Paper 3

Prediction of malt quality on whole grain and ground malt using near infrared spectroscopy and chemometrics

Jesper Pram Nielsen and Lars Munck

Prediction of malt quality on whole grain and ground malt using near infrared spectroscopy and chemometrics

Jesper Pram Nielsen and Lars Munck

Department of Dairy and Food Science, Food Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark. E-mail: jpn@kvl.dk.

Introduction

The quality of malt used in a brewery is important for the brewing process and the end quality. Quality evaluation of malting barley and malt is expensive and time-consuming, involving both micro-malting and mashing. In breeding, as well as in quality documentation, there is a need for fast and automated instrumental analyses. Near infrared (NIR) spectroscopy is a well-established rapid method for quality determination in cereals. The objective of the current investigation is to compare NIR transmission with NIR reflectance spectroscopy. In the NIR reflectance mode, measurements on whole malt grains are compared with measurements on ground samples and full spectra partial least squares models are compared with reduced models based on interval partial least squares (iPLS).[1]

Material and methods

50 micro-malt samples, representing 25 different varieties, grown at two different locations in Denmark, were analysed. The micro-malt samples are measured in three different modes: NIR transmission, Infratec 1255 Food and Feed Analyzer on whole grains in the range of 850–1048 nm; NIR reflectance, NIRSystems 6500 on whole grains; and NIR reflectance, NIRSystems on ground samples, both in the range of 400–2500 nm.

The samples are analysed for the five parameters: β-glucan in malt, nitrogen in malt, extract, modification and β-glucan in wort, all according to Analytica-EBC.[2] The chemometric calculations are performed using Unscrambler V. 7.01 (CAMO A/S, Trondheim, Norway) and Matlab V. 5.2 (The MathWorks, Inc.).

Results and discussion

The NIR transmission spectra are limited to the range of 850–1048 nm (100 datapoints) and are compared to the NIR reflectance spectra ranging from 400 to 2500 nm (1050 datapoints), which includes the visible spectral range. The NIR reflectance spectra of the 50 whole grain micro-malt samples are shown in Figure 1.

The NIR spectra are used for prediction of the five malt quality parameters using PLS regression. The 50 samples are divided into five subgroups of ten for validation. The models are compared according to their root mean squares error of cross-validation (*RMSECV*), by the correlation between measured and predicted and by the number of PLS components required.

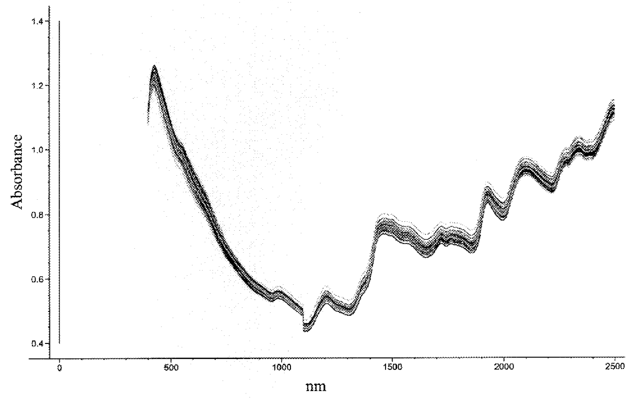


Figure 1. NIR reflectance spectra of whole malt kernels in the range of 400–2500 nm.

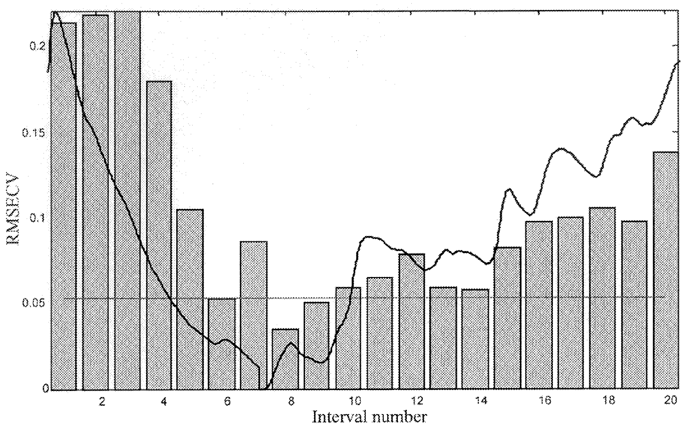


Figure 2. Interval selection using iPLS for the prediction of nitrogen in malt using the NIR reflectance spectra of whole malt kernels.

At Food Technology, The Royal Veterinary and Agricultural University, an iPLS algorithm has been developed[1] in which local PLS models of the full-spectrum are generated. In this way it is possible to focus on important spectral regions and remove interferences from other regions, thereby improving the model. After finding the region with the lowest *RMSECV*, the interval is further optimised by shifting the interval and changing the interval width.

Figure 2 shows an example in which NIR reflectance spectra of whole malt kernels are divided into 20 intervals. The *RMSECV* of the full-spectrum model predicting nitrogen in malt is shown, where the horizontal line indicates the *RMSECV* for the full-spectrum model, together with the average of the spectra. The bars represent the *RMSECV* for the different intervals and, as can be seen, interval number 8 has a considerably lower *RMSECV* than the full-spectrum model. Figure 3 shows the predicted v. measured plot, using the optimised interval ranging from 1130 to 1316 nm where the *RMSECV* is reduced from 0.05 to 0.03% N.

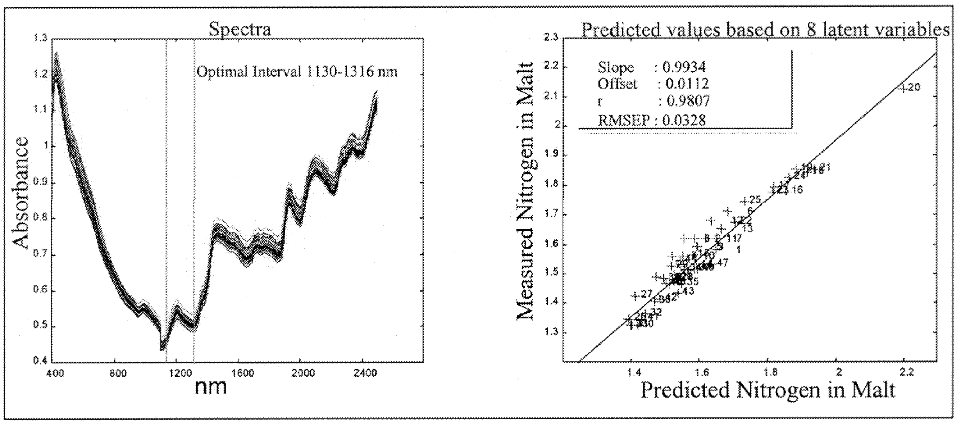


Figure 3. Prediction of nitrogen in malt using the spectral region of 1130–1316 nm.

Tables 1 and 2 summarise the performances of the NIR reflectance models, based on measurements on whole malt kernels and ground malt, with the full-spectrum models as well as with the models based on iPLS optimised models. In all ten NIR reflectance models, considerable improvements are seen, both with regard to prediction error and model complexity, i.e. the number of PLS components when using the optimised spectral region. No model improvements are seen when measuring ground sample compared to whole grain measurements. Transmission studies with the NIRSystems 6500 were unsuccessful (data not shown).

Table 1. Calibration results using NIR reflectance on whole malt grains. The full spectra models are compared with optimal iPLS interval. The table includes parameter, selected wavelength, correlation coefficient, number of PLS components (# PLS), root mean square error of cross-validation (*RMSECV*) and *RMSECV* divided by the range of the parameter (*RMSECV*/range).

Parameter	Wavelength (nm)	Correlation	# PLS	*RMSECV*	*RMSECV*/range (%)
β-glucan in malt	400–2500	0.89	10	0.20	12.6
	1330–1442	0.93	7	0.17	10.7
Nitrogen in malt	400–2500	0.95	10	0.05	6.3
	1130–1316	0.98	8	0.03	3.8
Extract	400–2500	0.92	9	0.6	8.8
	1204–1410	0.97	8	0.39	5.7
Modification	400–2500	0.87	10	5.6	12.4
	1348–1410	0.95	7	3.5	7.8
β-glucan in wort	400–2500	0.91	10	118	10.3
	1334–1436	0.97	7	75	6.5

Table 2. Calibration results using NIR reflectance on malt flour. The full spectra models are compared with optimal iPLS interval. The table includes parameter, selected wavelength, correlation coefficient, number of PLS components (#PLS), root mean square error of cross-validation (*RMSECV*) and *RMSECV* divided by the range of the parameter (*RMSECV*/range).

Parameter	Wavelength (nm)	Correlation	# PLS	*RMSECV*	RMSECV/range (%)
β-glucan in malt	400–2500	0.82	6	0.26	16.4
	1388–1598	0.87	4	0.22	13.9
Nitrogen in malt	400–2500	0.93	7	0.06	7.5
	2082–2164	0.95	2	0.05	6.3
Extract	400–2500	0.92	8	0.6	8.9
	2110–2140	0.94	3	0.48	7.0
Modification	400–2500	0.82	4	6.4	14.2
	2276–2386	0.87	2	5,4	12.0
β-glucan in wort	400–2500	0.92	10	112	9.8
	1368–1436	0.92	6	112	9.8

Table 3. Calibration results using NIR transmission on whole malt grains. The table includes parameter, selected wavelength, correlation coefficient, number of PLS components (#PLS), root mean square error of cross-validation (*RMSECV*) and *RMSECV* divided by the range of the parameter (*RMSECV*/range).

Parameter	Wavelength (nm)	Correlation	# PLS	*RMSECV*	*RMSECV*/range (%)
β-glucan in malt	850–1048	0.90	10	0.19	12.0
Nitrogen in malt	850–1048	0.97	6	0.04	5.0
Extract	850–1048	0.95	13	0.48	7.0
Modification	850–1048	0.89	10	5.3	11.7
β-glucan in wort	850–1048	0.91	10	126	11

For the NIR transmission Infratec spectra, no improvements were obtained using the iPLS algorithm, probably due to the narrow range. The performances of the full-spectrum NIR transmission models are shown in Table 3. Only minor differences in predictive performances are seen when comparing the optimised NIR reflectance models with the NIR transmission models, but the NIR reflectance models are considerably lower in model complexity.

Conclusions

NIR reflectance spectroscopy on whole malt grains can be used for determination of malt quality with accuracies comparable to near infrared transmission. The iPLS algorithm has improved NIR reflectance-based models considerably. The iPLS algorithm did not improve NIR transmission-based models.

References:

1. L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, submitted to *Appl. Spectrosc*. (1999).
2. *Analytica-EBC. Analysis by the European Brewery Convention*, 4. Edition. Brauerei- und Getränke-Rundschau, Zurich, Switzerland (1987).

Paper 4

Application of fuzzy logic and near-infrared spectroscopy for malt quality evaluation

Jesper Pram Nielsen, Rasmus Bro, Jørgen Larsen and Lars Munck

Abstract

An increasing number of quality criteria are involved in the evaluation of the final malt. This implies a comprehensive quality evaluation, normally based on experience and prior knowledge by the malster/brewer/breeder. This paper describes the principle in, and use of, fuzzy logic for the translation of a complex malt quality profile into a simple univariate overall quality index (OQI). The approach was tested on a data set of 50 malt samples including eleven quality parameters according to the European Brewery Convention.

The presented fuzzy logic approach involves three steps: i) an appropriate definition of how good a certain quality parameter level is, ii) a sound way to combine several quality parameters and iii) a way to express the overall quality based on all these individual parameters, taking their individual relative importance into account. The fuzzy logic based OQI presented here turned out to be a sound index for the overall quality of the tested malt samples, and thus provides a way of reducing and automating the quality data evaluation.

It is furthermore shown that near infrared transmittance spectra of the malt samples showed reasonable ability to predict the calculated OQI. Hereby, both analysis and evaluation efforts in malting barley breeding can be reduced considerably.

Introduction

Malting is basically a controlled germination of barley in which the starchy endosperm is modified into friable malt, ready for enzymatic degradation into fermentable sugars and amino acids to be used by the brewer's yeast. An increasing number of quality criteria are involved in the evaluation of the final malt. A complete quality analysis might contain 10-15 physical and chemical analyses, which makes the malt analysis costly and time consuming. Near infrared (NIR) spectroscopy has been used extensively for the prediction of a range of these malt quality parameters, as reviewed by Osborne *et al.* (1989) and Meurens and Yan (2002), and thereby contributed to a considerable reduction in the time of analysis. However, the high number of analyses, either determined by classical reference methods or by NIR spectroscopy, implies a comprehensive quality evaluation, as a given malt lot is to be evaluated on several quality parameters simultaneously. This evaluation is normally based on experience and prior knowledge by the malster/brewer/breeder, in which each quality parameter is evaluated according to a target or target range. The single value evaluations are then summarised in a total evaluation, to be used for final acceptance or rejection of a given malt sample.

The purpose of this investigation is to study the use of fuzzy logic for the translation of a complex malt quality profile into a simple univariate overall quality index. By using fuzzy logic it is possible to define in a simple way, i) an appropriate definition of how good a certain quality parameter level is, ii) a sound way to combine several quality parameters and iii) a way to express the total quality based on all these individual parameters taking their individual relative importance into account. At the heart of fuzzy logic is a membership function for each quality parameter. As an example of a membership function, consider the non-fuzzy logic implied when a person is asked to tell whether another person is young or old. The traditional binary logic implied requires that everyone beyond, say, 40 years is old (one), while everyone below is not (zero). It is apparent, however, that a person at 41 years is not much different with respect to age than a person at 39 years, even though the former would be considered old and the latter young. In a fuzzy system, the membership function enables age to be defined as a continuous function of age. For example, persons above 70 years may have a membership of one, meaning they belong completely to the class old. Persons at the age of 40 may have a membership of ½ meaning that these are equally young and old. Clearly, the fuzziness of the membership function is appropriate in this case and as will be shown, such fuzziness is also what is called for when quantifying the quality of malt with respect to different parameters. Fuzzy logic has found use in diverse disciplines since its formal introduction (Zadeh, 1965). There are several different types of fuzzy logic based inference systems (Jang and Sun, 1997; Mamdina and Assilian, 1975; Sugeno, 1985), and they have been applied in virtually all branches of applied science under slightly different names. Relevant reviews on fuzzy logic can be found in the literature (Saffiotti, 1997; Center, 1998). As the fuzzy system built in this paper is particularly simple, only the relevant theory will be described.

Several indices combining single barley and malt quality variables have been proposed. In the framework of the European Brewery Convention, a statistical index for the overall evaluation of malting and brewing quality in barley has been proposed (Molina-Cano *et al.*, 1986; Molina-Cano, 1987). This index is based on a weighted linear combination of extract, Kolbach Index, Apparent Final Attenuation, Viscosity and Diastatic Power. Monnez *et al.* (1987) proposed an index based on expert definition of barley groups based on their overall quality followed by linear discrimination between these groups.

The earlier approaches for calculating an overall quality index build on either a binary logic where each parameter is assessed compared to one specific target value or a continuous approach where the actual value of the parameter is used directly as being proportional to the quality (Molina-Cano *et al.*, 1986; Molina-Cano, 1987). These approaches are not satisfactory, since they unnaturally change the apparent expert knowledge into an overly restricted mathematical system.

For example, having one specific target value for a parameter is mostly not consistent with the a priori knowledge. For example, if a viscosity value of 1.10 is deemed optimal, then most likely a viscosity value of 1.20 or 1.40 is also optimal. Having the overall quality being linearly related to the level of a certain parameter throughout the parameter range is also not appropriate. Naturally, the relation between the level of a parameter and its evaluated effect on quality is highly nonlinear. Using the concept of fuzzy membership functions, will allow a simple way of treating the above problems. The method suggested in the following builds on membership function, and has no statistically estimated parameters. It is based on expert knowledge only, and therefore completely transparent from a user point-of-view.

Materials and Methods

Sample collection

The samples, provided by Carlsberg Research Laboratory, originate from trials harvested in 1995, under the European Brewery Convention (EBC). Fifteen spring barley varieties and ten winter barley varieties were grown at two different locations in Denmark, Jutland and Zealand, giving 50 malting barley samples in total. The barley grain samples were screened over a standard 2.5-mm sieve and the grains above 2.5 mm were subjected to the micro-malting and mashing procedure.

Malting quality analyses

The following 11 quality analyses were performed according to the official methods from the European Brewery Convention (EBC Analytica): ß-glucan in malt, extract, N in malt, viscosity, friability, malt modification, homogeneity, diastatic power, wort colour, soluble N in malt and ß-glucan in wort. These parameters are typical for a malt quality evaluation.

Near infrared transmittance (NIT) measurements

Near infrared transmittance spectra of the 50 malt samples (whole kernels) were recorded using an Infratec 1225 Food and Feed Analyzer (Foss Tecator, Höganäs, Sweden). The spectrophotometer records spectra in the range from 850 to 1050 nm with data collection at every 2 nm, yielding 100 data points as reported in absorbance (Log (1/T)). The whole malt kernels were loaded in a large vertical sample cell with a 30 mm path length and inserted into a transport module. Each spectrum is the average of 10 sub-scans acquired along the vertical sample cell.

Data analysis

The fuzzy logic calculations were performed using MATLAB version 6.1 (The MathWorks, Inc. Natick, MA) using the associated Fuzzy Logic toolbox. Principal Component Analysis (Wold *et al.*, 1987) and Partial Least Squares Regression (Martens and Næs, 1989) were computed using The Unscrambler 7-6 SR-1 (Camo A/S, Trondheim, Norway).

The principle of Fuzzy Logic

There are several steps in developing an overall quality index (OQI) using fuzzy logic. These steps are described in the following.

A simple non-binary (fuzzy) membership function is attached to each individual parameter, indicating to which degree any level of the parameter is good (with respect to malting). Thus, the membership function defines to which degree the quality parameter is acceptable and is defined on the basis of expert knowledge. Figure 1 shows an example of the membership function of soluble N in malt.

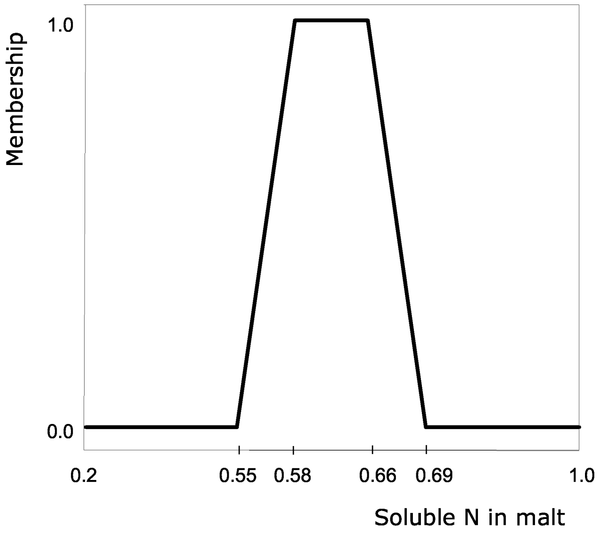


Figure 1. Example of a membership function of soluble N in malt

For any value of soluble N (x-axis), a corresponding membership can be read from the ordinate axis. For example, if soluble N is between 0.58 and 0.66, the membership will be one, indicating perfect malting quality (with respect to soluble N). Below 0.55, the membership, hence quality, is zero. Between 0.55 and 0.58 as well as between 0.66 and 0.69 is the interesting 'fuzzy' area, where the quality (membership) will increase or decrease. If the parameter lies within these areas, the malt is not unacceptable, though not optimal. The closer the parameter is to the optimal region, the better the malt is. This is quantified by the membership function which is a one-to-one mapping from the parameter space to membership space. The membership for any parameter is a number between zero (unacceptable) and one (optimal).

Membership functions can have virtually any shape (usually convex) but the exact shape is mostly not important. As long as the shape is in reasonable accordance with the background knowledge, reasonable results are obtained. For all 11 parameters in this investigation, the membership functions are defined on the basis

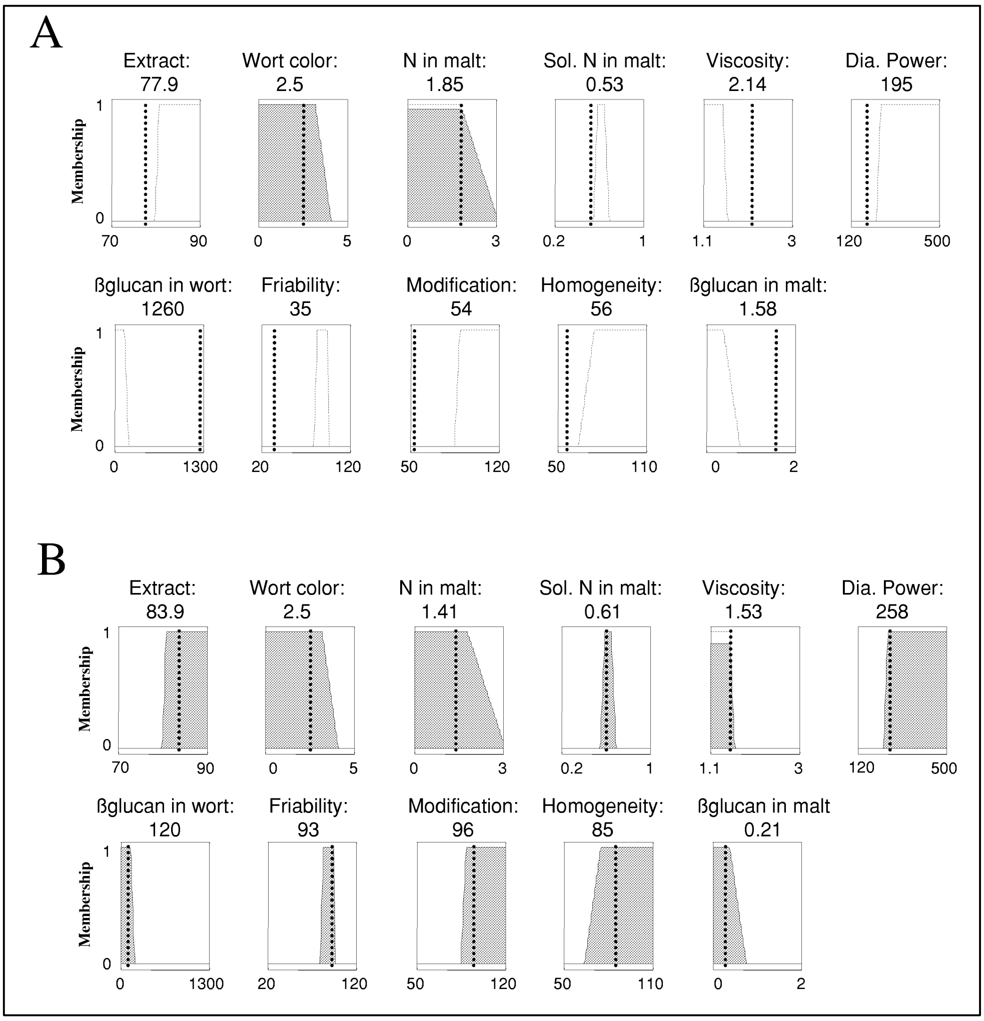


Figure 2. The membership functions for the 11 quality parameters. Two specific malt samples are also given. The top plots (A) show the least interesting malt sample and the lower (B) shows the best quality in the current sample set. For each parameter, the level of the parameter is defined on the x-axis and the membership (between zero and one) read on the y-axis.

of current subjective expert quality assessments and are shown in Figure 2. As can be seen, some parameters are simply better the higher (or lower), such as extract yield, while others have an optimal target region (e.g. soluble Nitrogen). This is easily handled by setting the shapes of the membership functions accordingly.When the membership functions are defined, any parameter can be converted into a membership degree for any malt sample. These memberships are then to be combined into one OQI.

The memberships (one for each quality parameter) are combined into the OQI by means of a simple weighted addition (sugeno-type). This type turned out to provide the most intuitive and directly appreciable quality measure. The weights are defined on the basis of expert evaluation of importance of each parameter where 10 is very important and 1 is less important. In Table I, the expert-defined weights are shown for the 11 parameters.

Table I. Importance weights of the 11 parameters

	Extract	Wort colour	N in malt	Sol. N in malt	Viscosity	Diastatic power	B-glucan in wort	Friability	Modification	Homoge-neity	B-glucan in malt
Weights	9	7	4	9	3	9	10	7	8	9	10

As for the definition of the membership functions, these weights are subjectively defined by the current evaluation of importance. OQI is then simply defined as

$$OQI = \sum_{i=1}^{11} m_i w_i$$

where m_{i} is the membership for parameter i 1 to 11 and w_{i} is the corresponding weights.

It should be emphasized that the membership functions as well as the importance weights have been chosen in order to fulfil current malt requirements based on the experience of the third author, but can be changed according to other requirements.

Results and Discussion

As an example, the fuzzy logic function described in the previous section was applied on a limited data set comprising of 50 malt samples representing 25 genotypes grown on two Danish locations. The detailed results of the eleven EBC malt analyses of the 50 samples used for this investigation are given in Table II.

Table II. Malt quality data of the 50 analysed samples including 15 spring barley and 10 winter barley varieties respectively grown in Jutland (number 1-25) and Zealand (number 26-50).

Number	Variety	Extract	Wort colour	N in malt	Sol. N in malt	Viscosity	Diastatic power	B-glucan in wort	Friability	Carlsberg Modification	Homoge-neity	B-glucan in malt
1	Alexis	81.7	2.2	1.58	0.55	1.64	254	260	83	88	58	0.42
2	Triumph	81.2	2.5	1.62	0.57	1.81	231	450	74	86	79	0.79
3	Nevada	80.9	1.9	1.59	0.51	1.96	232	580	72	81	62	0.84
4	Cooper	82.0	1.9	1.56	0.57	1.59	166	210	86	93	81	0.42
5	Caminant	81.4	3.0	1.59	0.63	1.85	201	420	72	82	65	0.84
6	Miralix	81.0	2.2	1.71	0.53	1.91	257	630	65	75	59	1.27
7	Texana	82.3	2.5	1.62	0.61	1.63	167	200	87	92	77	0.32
8	Trebon	81.9	2.5	1.62	0.58	1.59	307	150	77	92	60	0.26
9	Cork	81.0	2.2	1.49	0.50	1.57	373	180	80	91	59	0.32
10	Delibes	81.2	2.2	1.56	0.49	1.69	286	290	78	81	65	0.53
11	Polygena	82.4	2.2	1.62	0.60	1.60	356	240	84	91	78	0.32
12	Mentor	81.3	2.2	1.68	0.57	1.62	379	220	73	93	76	0.27
13	Mie	81.7	2.2	1.65	0.55	1.70	239	400	73	84	68	0.69
14	Reggae	82.4	2.5	1.53	0.53	1.68	215	400	79	84	70	0.74
15	Anni	81.0	2.5	1.57	0.50	2.16	239	1200	58	71	69	1.79
16	Plaisant	78.2	2.2	1.78	0.52	2.10	333	1200	38	51	68	1.79
17	Angora	79.6	2.5	1.80	0.61	1.73	378	770	48	71	53	1.15
18	Clarine	77.9	2.5	1.85	0.53	2.14	195	1260	35	55	56	1.58
19	Puffin	78.7	2.2	1.86	0.63	1.91	277	730	53	68	54	1.00
20	Geneva	77.1	2.5	2.13	0.70	1.83	437	650	44	71	66	1.00
21	Trasco	79.1	2.5	1.86	0.60	1.75	277	830	49	66	54	1.16
22	Fanfare	80.5	2.5	1.68	0.54	1.86	204	770	49	66	64	1.31
23	Melanie	79.9	2.5	1.78	0.60	1.77	388	750	58	71	56	1.15
24	Rejane	79.5	2.5	1.83	0.57	1.90	166	680	48	76	59	1.05
25	Sunrise	78.9	2.2	1.75	0.55	2.11	270	890	42	58	63	1.26
26	Alexis	82.7	2.5	1.52	0.59	1.56	305	180	88	93	71	0.26
27	Triumph	83.1	2.5	1.47	0.60	1.64	260	280	82	90	80	0.79
28	Nevada	82.1	2.5	1.45	0.51	1.78	238	450	79	84	64	0.63
29	Cooper	83.9	2.5	1.41	0.61	1.53	258	120	93	96	85	0.21

Continued on next page

Table II continued

30	Caminant	82.1	2.8	1.52	0.63	1.71	252	300	78	84	69	0.79
31	Miralix	83.4	2.8	1.50	0.61	1.60	158	180	89	94	75	0.26
32	Texana	82.7	2.5	1.49	0.55	1.67	146	500	80	87	77	0.68
33	Trebon	82.5	2.8	1.52	0.61	1.58	313	145	84	94	81	0.31
34	Cork	82.5	2.2	1.35	0.51	1.48	390	120	88	95	74	0.21
35	Delibes	82.6	2.7	1.42	0.50	1.56	287	150	89	95	82	0.21
36	Polygena	83.2	2.8	1.44	0.62	1.53	279	190	90	92	82	0.31
37	Mentor	82.6	2.8	1.55	0.63	1.57	316	150	82	92	72	0.31
38	Mie	83.1	2.8	1.47	0.58	1.56	212	220	86	90	71	0.31
39	Reggae	83.3	2.8	1.47	0.60	1.52	217	190	88	90	63	0.31
40	Anni	81.7	2.8	1.54	0.51	1.87	232	830	66	73	64	1.52
41	Plaisant	80.6	2.8	1.35	0.44	1.94	205	660	72	76	70	1.10
42	Angora	82.4	2.8	1.43	0.60	1.54	282	170	92	92	83	0.21
43	Clarine	81.0	3.1	1.49	0.47	1.87	151	700	68	73	64	0.94
44	Puffin	81.4	2.5	1.53	0.59	1.63	221	240	84	88	73	0.36
45	Geneva	80.9	2.5	1.71	0.63	1.65	384	260	81	85	72	0.42
46	Trasco	82.2	3.1	1.74	0.60	1.62	376	240	82	81	62	0.31
47	Fanfare	82.6	2.8	1.33	0.54	1.66	206	380	78	76	62	0.58
48	Melanie	82.7	3.1	1.33	0.61	1.54	314	170	92	92	76	0.21
49	Rejane	82.4	2.5	1.37	0.56	1.57	155	160	86	94	80	0.21
50	Sunrise	82.1	2.5	1.33	0.53	1.72	253	410	76	80	66	0.68

A normal evaluation of such a data table of malt quality results is based on experience and prior knowledge of the malster/brewer, in which each quality parameter is evaluated according to a target or target range. By applying the membership functions including the specific weights on to the original data (Table II) these data are converted from parameter space to weighted membership space. Thus, the original data matrix (50 samples x 11 variables) is converted into a new 50x11 matrix, where the variables represent new weighted optimality indices, one for each parameter. For instance, the first column (extract) has a weight of nine and is hence converted from the original range of 77.1 - 83.9 % extract yield to a weighted optimality index ranging from 0 (unacceptable) to 9 (optimal).

Exploration of this new weighted optimality matrix using PCA reveals five relatively distinct groups in a score plot of principal component (PC) 1 and PC 2 (Figure 3). These two principal components explain 70% of the variation, where PC 1 mainly explains differences in malt modification, while PC 2 mainly explains differences in diastatic power. The clusters can roughly be grouped as follows:

A:	not acceptable modification;	optimal diastatic power
B:	partly acceptable modification;	optimal diastatic power
C:	optimal modification;	optimal diastatic power
D:	not acceptable modification;	not acceptable diastatic power
E:	close to optimal modification;	not acceptable diastatic power

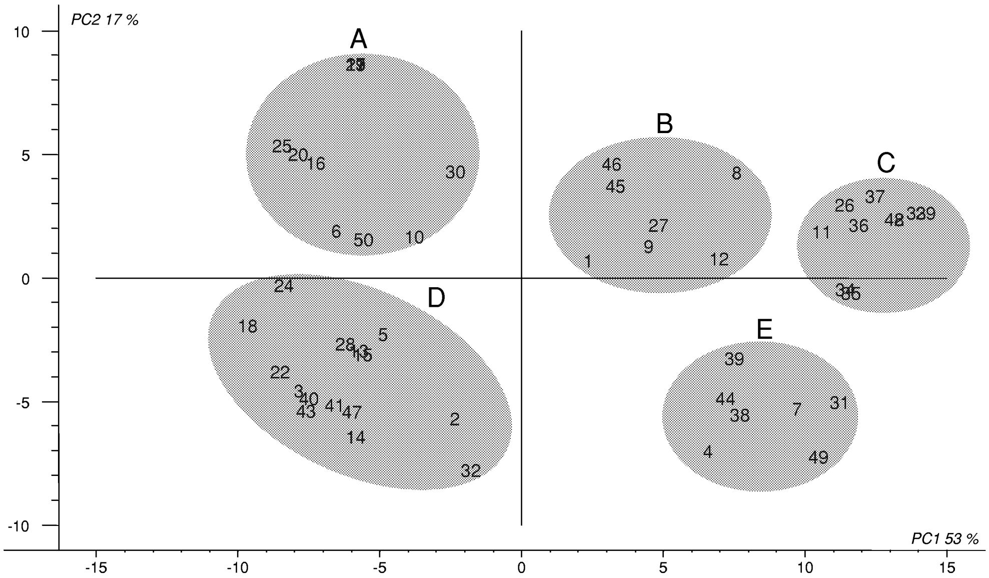


Figure 3. PCA score plot (PC 1 versus PC 2) of the weighted optimality matrix. The variety identification is given in Table II.

Performing a similar PCA analysis on the "raw" data in Table II (not shown) instead of the membership functions did not provide as clear results as the above, which is sensible considering that the two different representations provide different information. A PCA on the raw data focuses on the main directions in the original data, where the principal components are linear combinations of the original data and extreme samples will therefore have a large impact on the principal components. This is feasible for data overview, sample comparisons and interpretation of the principal components. However, a completely unsupervised PCA on the original data does not take into account the non-linear relationship between the parameter level and the degree of optimality as well as which levels are acceptable and which are not. Thus, by first converting the data into levels of optimality (weighted memberships) we supervise the PCA (Figure 3) to focus on differences in usability instead of the actual levels.

For each sample, the weighted optimality data are combined into a single overall quality index (OQI) by simple addition. An optimal sample would thus give an OQI of 85 (the sum of the weights multiplied by one), while a completely unacceptable sample would give an OQI of zero (the sum of the weights multiplied by zero). The OQI's of the 50 samples are given in Figure 4, showing considerable variations. Sample number 18 (Clarine grown in Jutland) seems to be the most unacceptable sample (OQI of 10.8), while sample number 29 (Cooper grown at Zealand) seems to be the most optimal (OQI of 84.7) of the analysed samples. The calculated OQI's of all the samples were thoroughly validated by the third author and found to be a sound index reflecting prior knowledge and expectations of the tested samples.

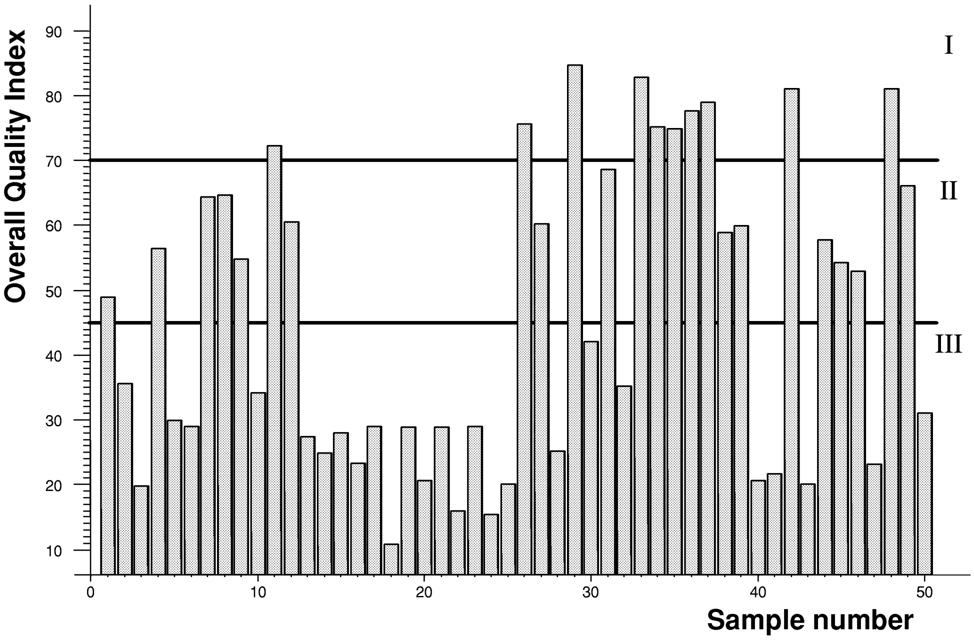


Figure 4. Bar plot of the calculated Overall Quality Index (OQI) of the 50 analysed malt samples. The variety identification is given in Table II. High overall quality index means good malt quality and low means poor. The horizontal lines indicate tentative limits for optimal (I), medium (II) and unacceptable (III) malt quality.

Two tentative limits have been added to Figure 4, indicating near optimal samples (I), medium samples (II) and not acceptable samples (III). A comparison between the PCA grouping in Figure 3 and the OQI's shows that the OQI group I is comprised of all the C samples in Figure 3, group II is comprised of B and E samples, while group III is comprised of A and D samples. Thus, the calculated

OQI does not clearly differentiate between B and E samples or between A and D samples. This shows that the calculated OQI as such only aims to determine whether a given sample is acceptable or not.

Combining the malt quality data into one single number facilitates plotting and easy data overview of the interactions between genotype and environment. Figure 5 shows the OQI's of the 25 genotypes grown on the two locations (x-axis Jutland and y-axis Zealand) including the tentative limits. It is evident that most of the samples perform considerably better in Zealand compared to Jutland (most genotypes located above the diagonal), while only Texana performs better in Jutland.

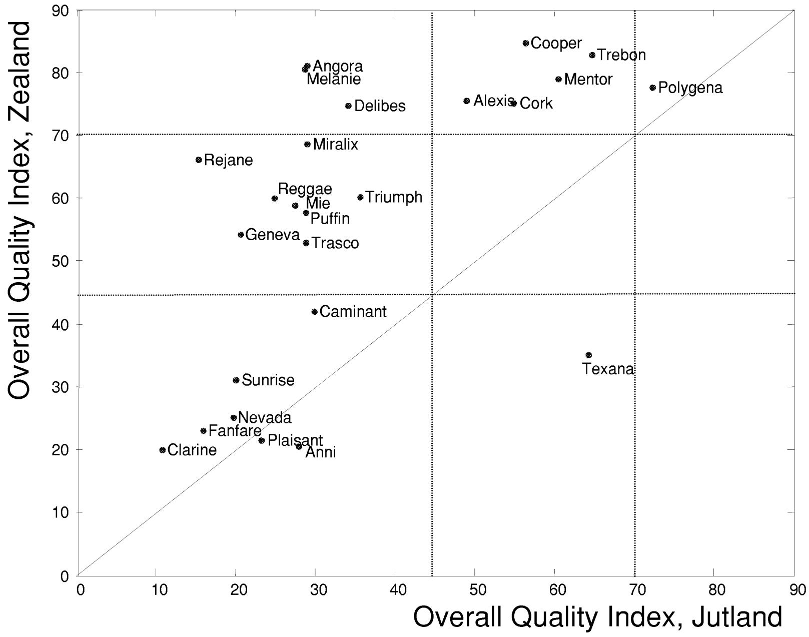


Figure 5. Scatter plot of OQI of the 25 genotypes grown on two locations (x-axis: Jutland; y-axis: Zealand)

Genotypes along the diagonal perform equally well on the two locations, while the non-diagonal samples are environmentally unstable (based on two locations only). As discussed earlier, Cooper grown in Zealand was the most optimal single sample. However, from a stability point of view, Polygena seems to perform similarly on both locations, even though on a slightly less optimal level. In the lower end, a group comprising Sunrise, Nevada, Fanfare, Plaisant, Anni and Clarine seems to be unacceptable on either location.

The fuzzy logic based OQI presented above is a way of reducing eleven malt quality parameters and thereby reducing the data evaluation efforts in, for example, malting barley breeding. However, the eleven input parameters are based on real malt quality analyses involving micro-malting and mashing. In order to reduce the analysis efforts as well, it would be of interest to predict this OQI by near-infrared spectroscopy. NIT spectra were recorded on the 50 malt samples, and the spectra were pre-transformed by the second derivative (Figure 6A) and used in a partial least squares regression (PLSR) model for the prediction of OQI.

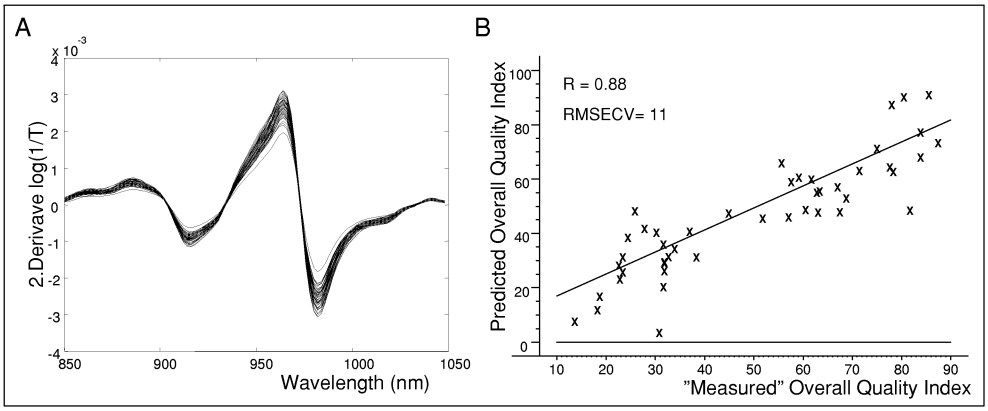


Figure 6. A) Second derivative NIT spectra in the range 850-1050 nm of the 50 analysed malt samples. B) Predicted versus “measured” OQI of a PLSR model using the second derivative NIT spectra.

Four outliers needed to be removed prior to modelling, namely the samples: 14, OQI=24.9; 16, OQI=23.3; 25, OQI=20.0 and 32, OQI=35.0 (See Table II for variety ID). These four samples are all in the lower end of the OQI scale, but the exact reason for these four being outliers remains unclear. The predicted versus measured plot of the remaining 46 samples is shown in Figure 6B, indicating a reasonable model with a correlation coefficient (r) of 0.88 and a cross-validated prediction error (RMSECV) of 11, which corresponds to approximately 15 % of the OQI range. This is not a perfect predictive model, but it is probably accurate enough for screening purposes in malting barley breeding by having the capability to classify the material into good, medium and bad categories. This will thus reduce both malt quality analyses and data evaluation.

Conclusions

This investigation has employed fuzzy logic for the translation of a complex malt quality profile into a simple univariate overall quality index. The approach was tested on a data set of 50 malt samples including eleven quality parameters. Fuzzy membership functions were constructed for each of the quality parameters. These functions define to which degree the quality parameter is acceptable on a scale from zero (unacceptable) to one (optimal), taking into account the non-binary and non-linear relationship between the level of the quality parameter and the degree of optimality. The memberships (one for each quality parameter) are combined into an overall quality index (OQI) by means of simple weighted addition. This OQI turned out to be a sound index for the overall quality. It is furthermore shown that a reasonable PLSR prediction model based on NIT spectra is obtainable. A relative prediction error of 15% was achieved, indicating a useful calibration for breeding purposes.

Acknowledgements

The authors are grateful to The Directorate for Food, Fisheries and Agri Business for financial support through the Danish Cereal Network-B4.

Literature

Analysis by the European Brewery Convention. Brauerei- und Getränke-Rundschau, Zurich (1987).

Center, B. 1998. Fuzzy Logic for Biological and Agricultural Systems. *Artificial Intelligence Review* 12 (1-3), 213-225.

Jang, J.-S. R. and Sun, C.-T. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, 1997.

Mamdani, E.H. and Assilian, S. 1975. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies* 7 (1), 1-13.

Martens H and Næs T. 1989. Multivariate Calibration. Wiley, New York.

Meurens, M. and Yan, S.H. Applications of Vibrational Spectroscopy in Brewing. In: Handbook of Vibrational Spectroscopy; Applications in Life, Pharmaceutical and Natural Science; Eds: Chalmers, J. M. and Griffiths, P.R. John Wiley and Sons, LTD, Chicester, UK, 2002.

Molina-Cano, J.L. 1987. The EBC Barley and Malt Committee Index for the Evaluation of Malting Quality in Barley and its Use in Breedning. Plant breeding 98, 249-256.

Molina-Cano, J.L., Madsen, B., Atherton, M. J., Drost, B.W., Larsen, J., Schildbach, R., Simiand, J.P. and Voglar, K. 1986. A Statistical Index for the Overall Evaluation of Malting and Brewing Quality in Barley. Monatsschrift für Brauwissenschaft 9, 328-335.

Monnez, J.M., Flayeux, R., Muller, P. and Moll, M. 1987. An approach to the estimation of brewing quality in barley and malt. Journal of the Institute of Brewing 93, 477-486.

Osborne, B., Fearn, T. and Hindle, P.H. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman Scientific and Technical, Harlow, UK (1993).

Saffiotti, A. 1997. The uses of fuzzy logic in autonomous robot navigation. *Soft Computing*. 1 (4), 180-197.

Sugeno, M. 1985 Industrial applications of fuzzy control, Elsevier Science Pub. Co.

Wold, S., K. Esbensen, and P. Geladi. 1987. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 37-52.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control* 8, 338-353.

Paper 5

Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics

Lars Munck, Jesper Pram Nielsen, Birthe Møller, Susanne Jacobsen,
Ib Søndergaard, Søren B. Engelsen, Lars Nørgaard and Rasmus Bro

ELSEVIER

Analytica Chimica Acta 446 (2001) 171–186

ANALYTICA
CHIMICA
ACTA

www.elsevier.com/locate/aca

Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics

L. Munck[a,*], J. Pram Nielsen[a], B. Møller[a], S. Jacobsen[b,1], I. Søndergaard[b,1], S.B. Engelsen[a], L. Nørgaard[a], R. Bro[a]

[a] *Chemometrics Group, Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

[b] *Section of Biochemistry and Nutrition, The Technical University of Denmark, Søltofts Plads, Building 224, DK-2800 Kgs. Lyngby, Denmark*

Received 7 November 2000; received in revised form 22 March 2001; accepted 4 April 2001

Abstract

Evaluating gene effects on proteomes and the resulting indirect pleiotropic effects through the cell machinery on the chemical phenotype constitutes a formidable challenge to the analytical chemist. This paper demonstrates that near-infrared (NIR) spectroscopy and chemometrics on the level of the barley seed phenotype is able to differentiate between genetic and environmental effects in a PCA model involving normal barley lines and the gene regulator *lys3a* in different genetic backgrounds. The gene drastically changes the proteome quantitatively and qualitatively, as displayed in two-dimensional electrophoresis, resulting in a radically changed amino acid and chemical composition. A synergy interval partial least squares regression model (si-PLSR) is tested to select combinations of spectral segments which have a high correlation to defined chemical components indicative of the *lys3a* gene, such as direct effects of the changed proteome, for example, the amide content, or indirect effects due to changes in carbohydrate and fat composition. It is concluded that the redundancy of biological information on the DNA sequence level is also represented at the phenotypic level in the dataset read by the NIR spectroscopic sensor from the chemical physical fingerprint. The PLS algorithm chooses spectral intervals which combine both direct and indirect proteome effects. This explains the robustness of NIR spectral predictions by PLSR for a wide range of chemical components. The new option of using spectroscopy, analytical chemistry and chemometrics in modeling the genetically based covariance of physical/chemical fingerprints of the intact phenotype in plant breeding and biotechnology is discussed.

Keywords: Proteome; Near-infrared spectroscopy; Chemometrics; Analytical chemistry; Biotechnology; Barley; Plant breeding

1. Introduction

Spectroscopy has become fundamental in chemistry. Its discovery dates back to 1800 when the British astronomer William Herschel reported on the existence of “the invisible thermometrical spectrum” to the Royal Society [1]. However, spectroscopy first gained momentum when Abney and Festing in 1881 [2] first measured spectra of organic compounds. Since then, a number of highly informative spectroscopic techniques have been developed. One of the more recent developments is near-infrared (NIR) spectroscopy which has invaded analytical chemistry by

[*] Corresponding author. Tel.: +45-35-28-33-58; fax: +45-35-28-32-45; URL: http://models.kvl.dk.
E-mail address: lmu@kvl.dk (L. Munck).
[1] URL: www.be.dtu.dk.

0003-2670/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0003-2670(01)01056-X

non-destructively predicting chemical composition, even in complex biological samples from agriculture [3]. These methods are now routinely used with high precision locally and internationally for controlling the composition of agricultural raw materials for feed and food such as cereals, milk, and meat.

It all started in the 1950s at the USDA laboratory in Beltsville, Maryland, USA, when Karl Norris and his group constructed a moisture metre [4] for wheat by NIR spectroscopy. The problem of interference in the moisture measurements from other chemical constituents was solved by data pretreatment through spectral derivatization and classical statistical wavelength selection by regression analysis made possible by the computer followed by multiple linear regression (MLR) to calibrate a few selected wavelengths to water measurements.

At the 7th World Cereal and Bread Congress in 1982, Martens and Jensen [5] introduced chemometric algorithms in utilizing information from whole NIR spectra. This was in the form of the partial least squares regression (PLSR) which later became fundamental in the development of software for NIR equipment dedicated for specific functions made by the instrumental industry.

In plant breeding, rapid screening methods for chemical composition and identification of specific genes are essential tools in classical breeding as well as in gene biotechnology. Recently, NIR spectroscopy has been shown to be able to detect the phenotypic effects of wheat–rye chromosomal translocations [6], and the chemical mechanism behind this classification was discussed.

In the 1960s, research in cereals was focused on obtaining genes for improved amino acid composition for nutritional purposes, especially with regard to the first limiting amino acid lysine [7]. The senior author of this paper was involved in developing a dye-binding method with acilane orange as an expression for the sum of the basic amino acids — lysine, arginine, and histidine — which was used as a ratio to protein (N × 6.25) to select the first *high-lysine* barley gene *lys1* from the world barley collection [8]. A more drastic ethylenimine-induced mutant M-1508 (gene *lys3a*) from the barley variety Bomi was isolated in 1973 by the Risø laboratory group [9] in Denmark employing the dye-binding method. The regulatory status of the Mendelian high-lysine gene *lys3a* was finally established in 1996 [10]. In a recent study, we have demonstrated that NIR spectroscopy is able to differentiate between five different high-lysine mutant genotypes [11] with characteristically different amino acid patterns.

In an autopollinated crop such as barley, spectroscopic screening is greatly facilitated analytically, because each line derived after six to nine generations of self-pollination can be considered homozygotic and thus genetically homogeneous. A mutation or a transfer of a specific gene to such a line is thus expressed in a genetically reproducible, controlled isogenic background and will show up in the spectroscopic physical/chemical fingerprint, if its chemical implications directly or indirectly are large enough. In development of NIR analytical methods, the applications have always been ahead of theory. We therefore aim at exploring the chemical basis of how NIR spectroscopy works in differentiating the *lys3a* phenotype from normal barley in two different environments. The genetically based diversity is first detected on the phenotypic level of biological organization by non-invasive spectroscopy and afterwards calibrated to destructive chemical analytical methods. This dialogue between data from the biological and chemical levels of organization is made possible by chemometric software and the computer.

2. Materials and methods

Total 125 different varieties of normal barley (O) and *lys3a* (X) lines based on crosses with these varieties were bred at the Carlsberg Research Laboratory, Valby, Copenhagen from 1973 to 1990, as represented in the figures and tables. They were grown in the field and/or in the greenhouse (V) together with the original *lys3a* mutant from Risø (M-1508) and its isogenic motherline Bomi. Seeds grown in the greenhouse tended to have a lighter color and higher protein content compared to those grown in the field. The whole seed barley samples were measured with a near-infrared transmission (NIT) instrument, Infratec (Foss Tecator AB, Høganäs, Sweden). The samples were milled in a hammermill (sieve 0.5 mm) and the whole flour was measured by a NIR instrument (Foss-NIR-Systems 6500, USA). The samples were analyzed for moisture, Kjeldahl protein (N × 6.25)

(Foss Tecator, Kjeltec) and for amides by alkali volatile nitrogen by adding 50 ml of 36% NOH alkali to 3 g of flour in the Kjeldahl destillation unit.

Twenty-one of these samples presented in figures and tables were also analyzed for starch (AACC 76-13), for fat (Foss Tecator, Soxtech), for β-glucan (Foss Tecator β-glucan analyzer system, Carlsberg), for soluble and insoluble fiber (Foss Tecator, Fibertech), and for amino acids after hydrolysis [12]. The 21 barley samples were extracted in order to obtain buffer-soluble (albumins, globulins) and ethanol-soluble (hordeins) proteins which were separated by a two-dimensional gel electrophoresis [13]. Principal component analysis (PCA) and PLSR analyses were performed by the "Unscrambler" software version 7.5 (Camo A/S Trondheim, Norway) with full cross-validation. The spectroscopic data were reduced by 50% by selecting information from every second wavelength with full cross-validation.

Interval-PLS (i-PLS) employed in the selection of wavelength areas [14] was performed on the 1050 NIR spectral data points divided into 30 equal intervals numbered 1–30, stating correlation coefficients (*r*) and error (RMSECV: root mean square error of cross-validation). An extension of i-PLS called synergy i-PLS (si-PLS) was employed to find the interval combinations of all possible combinations of intervals which give the highest correlation coefficients and the lowest errors.

3. Results from observations and experiments

3.1. Exploratory classification of separate spectral and chemical datasets by PCA

We will first investigate the non-destructive observation of batches of whole barley seeds by NIT spectroscopy and its ability to differentiate between *lys3a* and normal phenotypes. In Fig. 1A, second derivative NIT spectra between 860 and 1035 nm from 51 barley samples grown in the field are shown. We will now simulate the discovery of the *lys3a* gene, as it could have taken place with NIT spectroscopy instead of the dye-binding method [6–8]. In Fig. 1B, a PCA of 51 barley NIT spectra displays a normal barley population (O) and a *lys3a* mutant outlier (X). When this mutant is crossed with different normal barley genotypes, the segregants form two clusters (Fig. 1C) representing normal barley (O) and *lys3a* mutant (X) recombinants.

It is obvious that the non-destructive NIT spectroscopy on whole seeds is an attractive method for selection in plant breeding and that it is able to differentiate between the two extreme genotypes with a sufficient degree of precision. However, the short spectral range of 175 nm of NIT is mainly due to the third and fourth and partly the second overtones in the lower range of the near-infrared spectrum as limited by the silicon sensor. Further development of NIT spectroscopy with new sensors will reveal how far this technology can be expanded upwards at higher wavelengths to obtain less crude and more detailed spectra as with near-infrared reflection (compare Fig. 1A with D). Expanding upwards in the near-infrared spectrum, will give more specific chemical information, including from the combinatory region from 1900 nm and upwards [15]. Therefore, in order to explore these possibilities, we have chosen in the following to concentrate on NIR spectroscopy with photomultipliers measuring 1050 data points at every second wavelength from 400 to 2500 nm (Fig. 1D). This however introduces the drawback of having to mill the barley seed samples.

In Fig. 2 in a PCA plot with 125 NIR spectra constituting the whole barley material, we can identify four clusters where PC 1 differentiates between the genotypes normal (O) and *lys3a* (X) barley, while PC 2 differentiates between the barley grown in the field (O and X) and in the greenhouse (OV and XV). With few exceptions, the genetic differentiation is excellent. It would probably have been even better, if samples of the mutant had been compared with several samples of a normal barley with the same isogenic background.

Total 21 of these samples, 15 normal and 6 *lys3a* lines, were subjected to a detailed chemical analysis, including two two-dimensional gel electrophoresis analyses with a buffer and an ethanol extract of each sample in order to study the water- and salt-soluble albumins and globulins as well as the ethanol-soluble storage proteins, the prolamins (hordeins). By visual inspection, it was clearly possible to classify the two electrophoresis patterns of each of the pure samples of the *lys3a* genotype as different from those of the normal barleys. Two representative sets of electrophoresis, each for water/salt-soluble and ethanol-soluble proteins for the original 1508 gene a *lys3a* and the

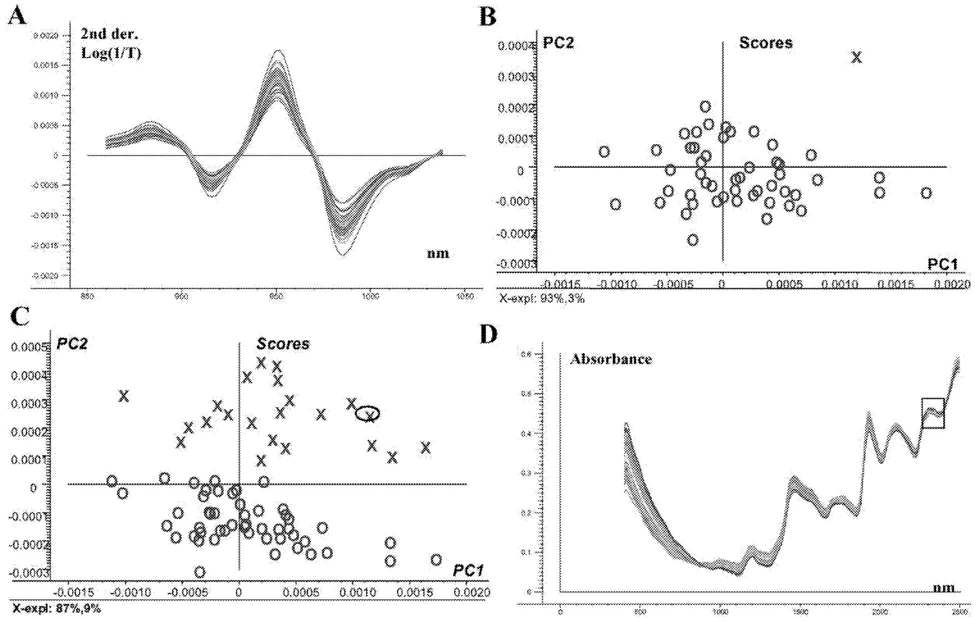


Fig. 1. (A) NIT spectra (850–1040 nm) measured non-destructively on whole seed samples of normal and *lys3a* barleys grown in the field presented as PCA plots in (B–D); (B) PCA of whole seed NIT spectra detecting a *lys3a* (mutant 1508) outlier (X) among normal barleys (O); (C) PCA of whole seed NIT spectra showing a segregating population for the *lys3a* gene (X) where the original 1508 mutant (encircled) has been crossed with normal barley (O); (D) 125 NIR spectra (400–2500 nm) from whole flour of milled seeds of normal and *lys3a* barleys shown in a PCA plot in Fig. 2. The squared area 2270–2360 nm (approximately identical with the i-PLS interval number 28) visually selected for difference between normal and *lys3a* spectra is presented enlarged in Fig. 6.

isogenic barley line Bomi are displayed in Fig. 3. The different proteomes give a good, general overview of the very high chemical complexity on the protein level which fuels the phenotypical variation discussed above.

The hordein patterns from a normal barley variety such as Bomi compared with those of the high-lysine mutant *lys3a* in Bomi shows that the normal line expresses a number of proteins that the mutant does not express (Fig. 3a and b). Contrary to this finding, the albumin/globulin fraction shows that the mutant line expresses a number of proteins that Bomi does not express (Fig. 3c and d). In the albumin/globulin fraction, the picture is rather complicated. A general conclusion about whether these differences are of a quantitative or a qualitative nature merits further studies during which the electrophoresis separation is optimized.

The changes in the proteome of the barley endosperm due to the *lys3a* gene are reflected in a drastic change in the amino acid composition of the total protein (see Table 1). The basic amino acids, including lysine, are increased together with aspargine, alanine, threonine, and valine, while glutamic acid, proline, amide nitrogen to protein nitrogen (A/P index) (see Table 2) and phenylalanine are markedly decreased. The change in the amino acid pattern of the 21 samples is expressed in the PCA biplot with scores and loadings in Fig. 4. There is a clear differentiation between *lys3a* (X) and normal barley samples (O). However, two of the samples, X and XV, are more intermediate, emerging from the same 508 line of the

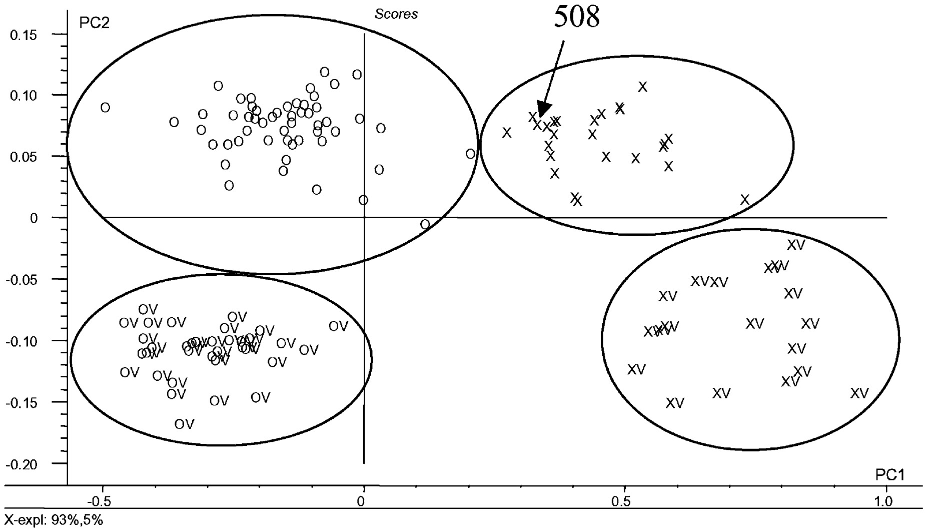


Fig. 2. PCA on NIR spectra from Fig. 1D from normal (O) and *lys3a* (OX) barleys grown in the field and in the greenhouse, OV and XV, respectively.

Carlsberg breeding program. The loadings from the different amino acids appear in three groups. On the right side of the PCA plot near the *lys3a* barleys (X and XV), there is a cluster of amino acids such as lysine and aspartic acid which are increased in this genotype. Situated above this are the amino acids with minor changes between the two genotypes, while to the left near to the normal (O and OV) samples are the amino acids like glutamic acid, proline, and A/P which are high in the normal genotypes. It is seen that the normal barleys grown in the greenhouse (OV) are near this amino acid cluster, because they have higher amounts of the amino acids typical for storage proteins like glutamic acid and proline compared to

Table 1
Selected barleys for amino acid determination (g per 100 g protein) from the material in Fig. 2

	Normal barley (n = 15)	Barley 1508 (n = 4)
Asp	4.78 ± 0.50	8.11 ± 0.57
Arg	4.51 ± 0.50	6.97 ± 0.42
Lys	3.10 ± 0.39	4.84 ± 0.28
Gly	2.93 ± 0.33	4.30 ± 0.32
His	v2.12 ± 0.21	2.76 ± 0.14
Tyr	2.81 ± 0.26	2.93 ± 0.22
Ala	3.07 ± 0.36	4.23 ± 0.28
Ser	3.17 ± 0.28	3.49 ± 0.33
Thr	2.63 ± 0.27	3.28 ± 0.30
Val	4.32 ± 0.45	4.92 ± 0.37
Met	1.43 ± 0.15	1.53 ± 0.14
Cys	1.56 ± 0.28	1.60 ± 0.41
Leu	5.98 ± 0.59	5.85 ± 0.49
Ile	3.12 ± 0.28	2.99 ± 0.19
Glu	21.62 ± 2.51	14.60 ± 1.69
Pro	9.63 ± 1.34	5.76 ± 0.64
Phe	4.51 ± 0.50	3.67 ± 0.22

Table 2
Chemical composition (% DM) of barleys for the same material selected for amino acid composition in Table 1

	Normal barley	Barley 1508
	(n = 15)	(n = 4)
Protein (N × 6.25)	12.76 ± 2.38	14.28 ± 1.84
Amide-N	0.32 ± 0.08	0.25 ± 0.05
Amide-N/N (A/P)	15.68 ± 1.08	11.53 ± 1.98
Beta-glucan	4.76 ± 0.77	3.88 ± 1.19
Fat	1.88 ± 0.18	3.22 ± 0.53
Starch	54.83 ± 4.18	50.95 ± 3.03
Insoluble fiber	10.91 ± 1.64	16.47 ± 0.91
Soluble fiber	2.87 ± 0.67	2.30 ± 0.69

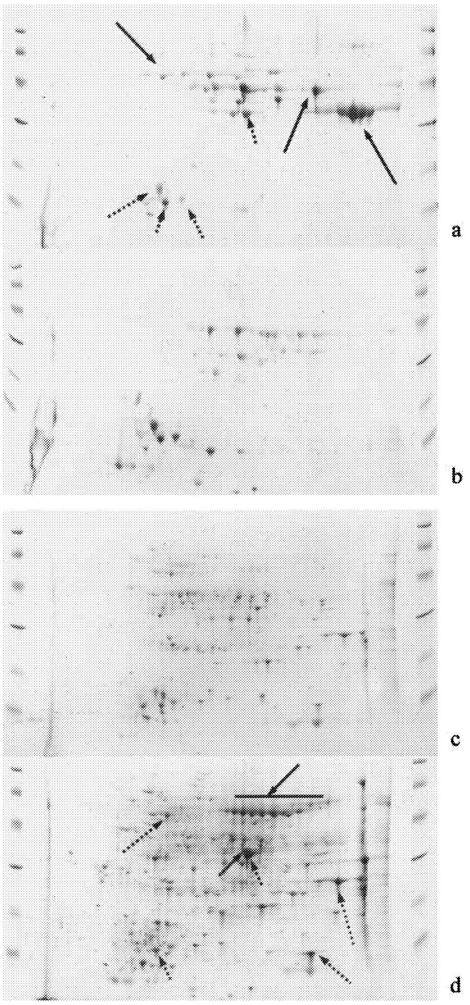


Fig. 3. Two-dimensional electrophoretic gels [13] of hordeins and albumins/globulins from the barley variety Bomi and its *lys3a* mutant. The gels were run using an immobilized pH gradient from 3 to 10 in the first dimension: (a) hordeins from Bomi; (b) hordeins from its *lys3a* mutant; (c) albumins/globulins from Bomi; (d) the albumins/globulins from its *lys3a* mutant. Some protein spots can be seen in the gels both from Bomi and the mutant, a subset of these common spots being indicated with dashed arrows for orientation. The full arrows indicate a subset of proteins that are only present in either Bomi or *lys3a*.

those grown in the field (O). This is caused by the higher protein content of the samples grown in the greenhouse (V) due to intensive nitrogen fertilization which especially increases the alcohol-soluble storage proteins, the hordeins.

Simply inherited Mendelian regulating genes like *lys3a* have complicated, indirect effects on the phenotype when the changed protein pattern influences the total endosperm cell machinery. This is reflected in a change in the total chemical composition with increases in *lys3a* barley for protein, fat and insoluble fiber and decreases in amide-N, beta-glucan, starch, and soluble fiber (Table 2 and Fig. 5). The PCA biplot of the total chemical composition in Fig. 5 parallels that of amino acid composition in Fig. 4, elucidating the different patterns in chemical composition due to genotype and growth environment. Also here, two of the *lys3a* lines, X508 and XV508, are intermediate. A close inspection of the seeds facilitated by the fact that *lys3a* seeds have a large embryo [7] reveals that these lines are not pure, but contain 39% (X508) and 60% (XV508) normal barley seeds on weight basis. This impurity is to some extent reflected in the position of the X508 sample adjacent to the normals (O) in the PCA plot of the whole material in Fig. 2; the spectrum of the XV508 sample was not included.

3.2. Establishing causal relationships between spectral, genetic, and chemical information by PLSR and wavelength selection

In spectroscopic evaluation, it is important from the onset of the investigation to carefully inspect the individual spectra. Inspired by the spectral variation, a trained NIR spectroscopist is able to select several wavelengths which may contribute to the chemical validation of the problem. As an example of visual selection, we will discuss a small spectral area displaying an interesting, fine structure between 2270 and 2360 nm marked by a square in Fig. 1D and enlarged about 20 times in Fig. 6A and B. In Fig. 6A, we can compare the spectrum of the original M-1508 mutant with that of its isogenic motherline Bomi, displaying two entirely different patterns. The Bomi spectrum compared to that of *lys3a* shows a more marked shoulder from about 2283 to 2295 nm and a maximum (instead of a decrease) at about 2320 nm, while the *lys3a* spectrum has a dual peak at about 2315 and 2345 nm.

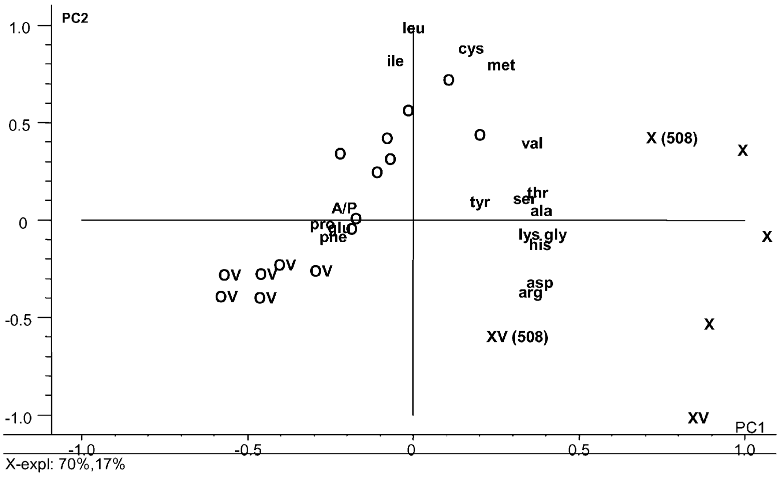


Fig. 4. A PCA biplot with scores and loadings of 21 amino acid analyses of normal (O) and *lys3a* (X) barley samples (see Table 1). V denotes barleys grown in the field; 508 is a putative *lys3a* line which is an outlier due to a contamination with normal barley seeds. The amino acid symbols denote loadings. The A/P symbol denotes loadings for the amide-N to total N index.

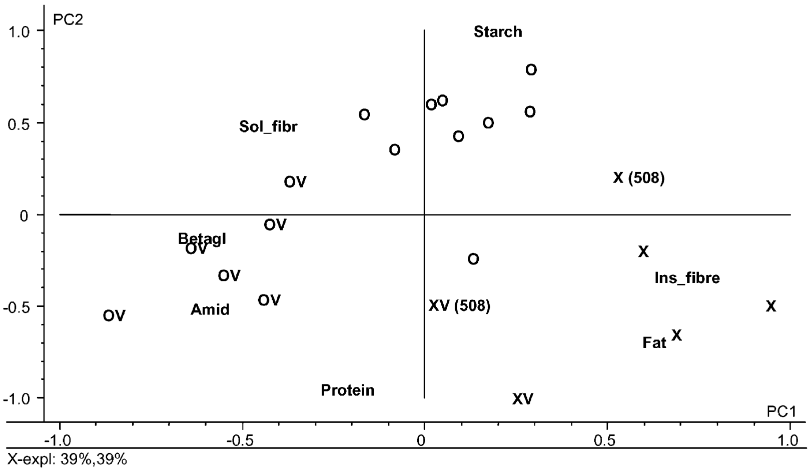


Fig. 5. A PCA biplot with scores and loadings of the chemical composition of the 21 normal (O) and *lys3a* (X) barleys from Table 2. V denotes barleys grown in the field; 508 is a *lys3a* outlier described in Fig. 4.

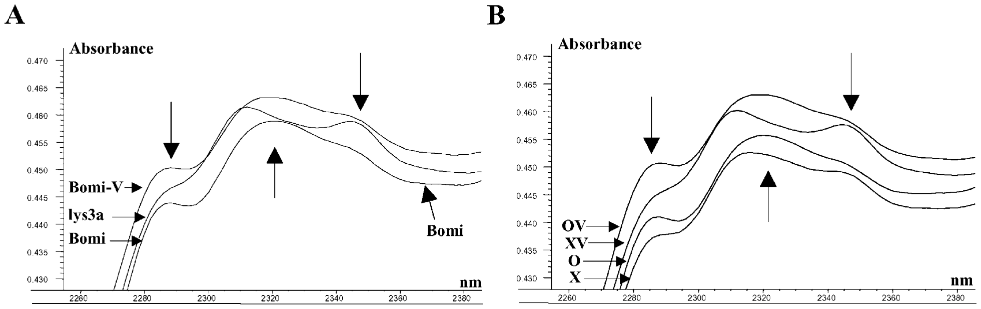


Fig. 6. Enlargement of part of the NIR spectrum 2290–2360 nm in Fig. 1D marked with a square. The arrows denote differences between the spectra which enable a classification discussed in the text — Fig. 1A: spectra of *lys3a* and its isogenic motherline Bomi grown in the field together with that of Bomi grown in the greenhouse (V); Fig. 1B: mean NIR spectra of normal barley (O, $n = 52$; OV, $n = 30$) and *lys3a* barley (X, $n = 25$; XV, $n = 18$).

Both these spectra represent samples grown in the field.

In comparison, Bomi-V (Fig. 6A) grown in the greenhouse (V) demonstrates essentially the same spectral form as Bomi grown in the field, but at higher absorbance reflecting the higher protein content. The conclusions from Fig. 6A regarding the isogenic lines and locations are confirmed for the barley varieties (O, OV) and *lys3a* crosses (X, XV) in Fig. 6B, displaying the mean spectra of the four classes: O ($n = 52$), OV ($n = 30$), X ($n = 25$), and XV ($n = 18$). Returning to Fig. 1D for comparison, an impressive reproducibility of the NIR spectral measurements is demonstrated. We may conclude that by visual inspection of the region 2270–2360 nm, it is possible to correctly classify spectra from normal and *lys3a* barley. The two contaminated deviating *lys3a* lines marked 508 discussed above show spectral characteristics intermediate between *lys3a* and the wildtype. Upon consulting the spectral table for chemical assignment [15], it appears that at 2294 nm there is an amino acid determinant (N–H and C=O) at the normal barley plateau of 2283–2295 nm. The normal barley peak at 2320 nm does not seem to coincide with any nitrogen bond information, but rather with CH information at 2310 nm (CH_2) and 2323 nm (CH_2). At 2336 nm, there is information on cellulose. The *lys3a* peak at 2345 nm is close to the $HC{=}CHCH_2$ indication at 2347 nm for unsaturated fat. It can thus be concluded that the small area of 90 nm between 2270 and 2360 nm, apparently unique for the *lys3a* genotype, is characterized not only by differences to normal barley in amino acids (protein), but also in carbohydrates (cellulose) and (unsaturated) fat, as confirmed in the chemical analyses (Table 2 and Fig. 5). These differences are to be related to the part of the proteome regulated by the *lys3a* gene which directs the machinery of the endosperm and germ tissue cells during development.

We will now supplement wavelength characterization by the naked eye with automatic selection using chemometric algorithms for data reduction into latent factors (principal components), such as with PLSR [5] from spectral intervals created by i-PLSR [12]. In order to indicate which wavelengths are dependant on genotype (normal and *lys3a*) and location (field and greenhouse) in the total material from Fig. 2, a discriminant PLSR was made with a 2 × 2 factor setup (wildtype = 1, *lys3a* = 0, field = 1, greenhouse = 0). The regression coefficients related to spectral wavelengths are presented in Fig. 7. The genotype component has a much higher profile than that of the location with positive and negative peaks at 495, 510, 1040, 1375, 1505, 1650, 1890, 1900, and 2400 nm. There were however large unique areas in the location loading which explained the satisfactory PCA classification in Fig. 2.

In further defining the chemistry behind the NIR spectroscopy of the barley material, we will now use PLSR to calibrate spectroscopic information on the level of the seed phenotype with two basic chemical

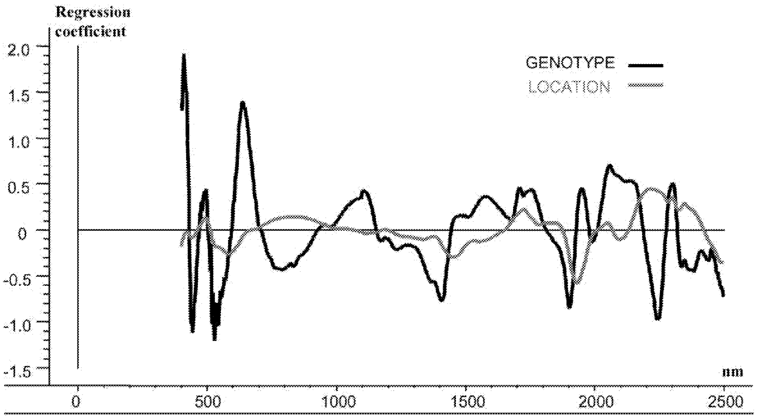


Fig. 7. Regression coefficient from a discriminant PLSR involving NIR spectra, x (n = 125) and genotype (normal and *lys3a* barley) and location (grown in greenhouse and in the field) y, from the data in Fig. 2.

analyses, Kjeldahl protein equal to N × 6.25 which represents univariate expression of a range of nitrogen bonds, and the more specific alkali volatile nitrogen analysis which mainly represents the amide-N group and small amounts of ammonium salts available in the barley seeds. In Fig. 8, high correlation PLSR models (six to seven PLS components) with full cross-validation are calculated based on whole spectra for prediction of protein (Fig. 8A), amide-N (Fig. 8B), and the ratio between amide-N and total N (Fig. 8C). It is seen that the barleys grown in the greenhouse marked V tend to have the highest amount of protein (Fig. 8A) and amide-N (Fig. 8B) content and that the *lys3a* genotype marked X generally has a lower content of amide-N compared to normal barley (O). This tendency is further reinforced in the PLSR model for prediction of the amide-N/total N ratio where there is a clear-cut clustering which separates the *lys3a* (X) genotypes from the normal (O). The amide-N/total N ratio available as spectroscopic information is thus one of the many spectral methods effective for screening for the *lys3a* genotype as an alternative to the dye-binding method [7].

In order to further dissect and explain the spectral information, the spectra were divided into 30 intervals of 70 nm each, giving 35 data points after 50% reduction. These were calibrated to protein and amide-N in fully cross-validated i-PLS [14] models. An example with the distribution of RMSECV along the spectral intervals is shown for amide-N in Fig. 8D for the whole barley material (n = 125) where the spectral interval number 23 (1940–2008) shows the lowest error. The covariance is generally high and the correlation coefficients with, for example, amide-N content vary typically between 0.99 (highest) and 0.80 (lowest) for five principal components for the 30 intervals.

In order to study synergy between the different spectral intervals, a si-PLS model was developed for computational time reasons limited here to two combined segments. Fig. 8E shows the two selected spectral intervals for amide-N, numbers 23 and 26, and the PLSR model is displayed in Fig. 8F.

For comparison of interval selection by si-PLSR, three data materials were constructed: normal barley (n = 82), *lys3a* barley (n = 43), and the total barley material normal + *lys3a* (n = 125). They were analyzed for si-PLS calibrated to protein and amide-N with full cross-validation. In Table 3, the intervals that were selected with maximum correlation coefficients are presented and the number of PLS components noted. The correlation coefficients and errors for the adjacent principal components were also selected and compared to a full-spectrum PLSR model. It is seen that the information is widely confounded and

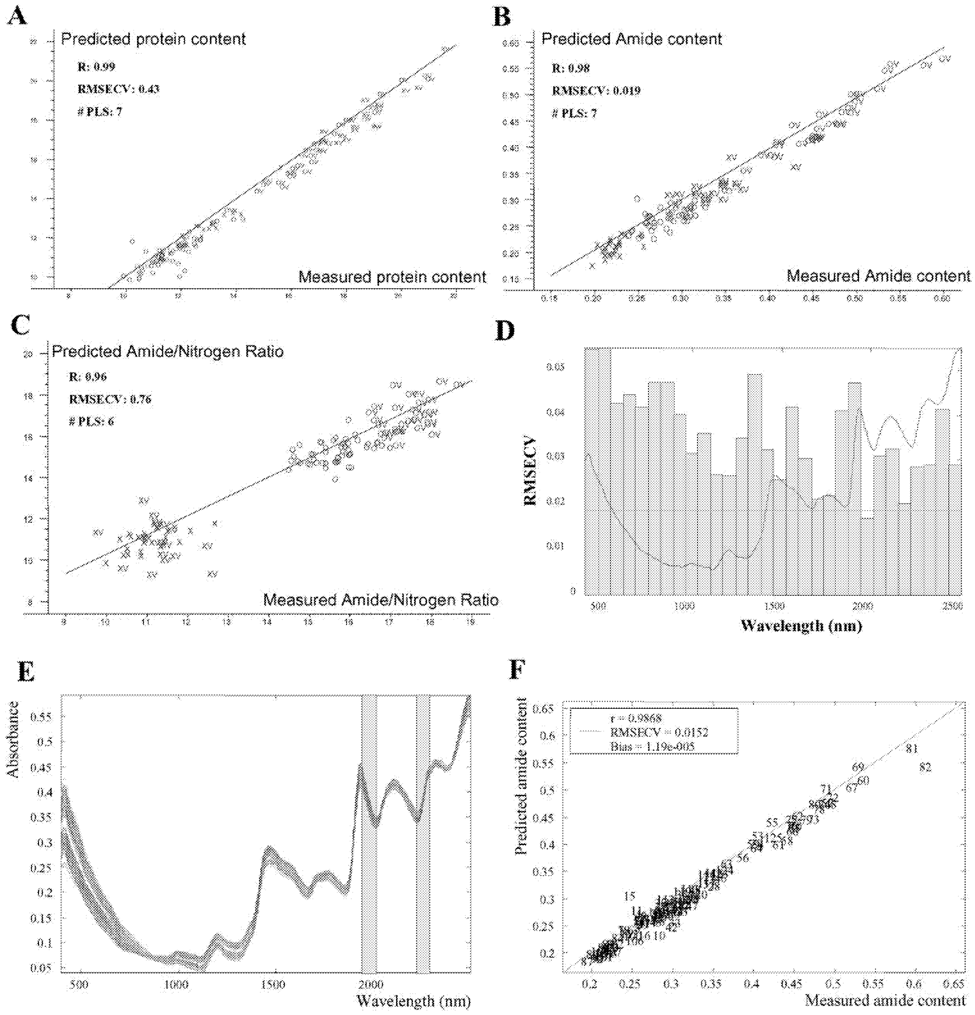


Fig. 8. PLSR regressions of the NIR material in Fig. 2 ($n = 125$) representing prediction of protein (N × 6.25) in (A), of amide in (B) and of the amide-N to total nitrogen ratio in (C). The RMSECVs for the prediction of amide content from NIR spectra of the 30 wavelength intervals chosen for the i-PLSR are shown in (D) and the two optimized wavelength intervals for amide prediction in (E), whereas (F) denotes the PLSR calculation for these two intervals.

Table 3
Selection of NIR spectral segments from different barley materials by si-PLSR calibrated to protein N and amide-N[a]

Calibration	Material											
	Normal (O) + *lys3a* (X) (n = 125)				*lys3a* (X) (n = 43)				Normal (O) (n = 82)			
	PLS components	Intervals	RMSECV	r	PLS components	Intervals	RMSECV	r	PLS components	Intervals	RMSECV	r
Protein N × 6.25	6	16–18	0.42	0.991	5	15–26	0.25	0.997	5	25–27	0.46	0.988
	7	20–27	0.39	0.992	6	20–27	0.25	0.997	6	24–27	0.43	0.989
	8	16–18	0.40	0.992	7	20–27	0.26	0.997	7	15–26	0.44	0.988
	7	FS[b]	0.43	0.990	7	FS	0.32	0.995	7	FS	0.46	0.988
Amide-N	7	12–23	0.015	0.987	6	25–26	0.0087	0.989	6	20–26	0.0161	0.986
	8	23–27	0.015	0.987	7	24–26	0.0076	0.992	7	20–26	0.0158	0.986
	9	15–23	0.014	0.988	8	23–27	0.0077	0.992	8	15–26	0.0162	0.986
	7	FS	0.019	0.980	7	FS	0.0196	0.944	7	FS	0.0167	0.985

[a] Interval synergi-PLSR selection for the barley NIR spectra; 30 spectral intervals, two synergetic components.
[b] Full spectrum.

Table 4
Chemical characterization [15] of wavelength intervals selected by si-PLSR in Table 3

Interval number	Selected in correlation with[a]	Wavelength interval (nm)	Chemistry [15]				
12	A	1170–1238				CH	
15	P, A	1380–1448	NH		ROH	CH	
16	P	1450–1518	NH, protein		ROH, starch cellulose		
18	P	1590–1658				CH	
20	P, A	1730–1798			Cellulose	CH	SH
23	P, A	1940–2008	NH	Amide	R–OH, starch		
24	P, A	2010–2078	NH, protein	Amide, C–O			
25	P, A	2080–2148		Amide	Starch, C=C (fat)		
26	P, A	2150–2218		Amide	Fat, CHO		
27	P, A	2220–2288	Amino acid		Starch	CH	
28[b]		2290–2358	Amino acid		Cellulose, fat	CH	

[a] P: protein; A: amide.

[b] Visually selected.

there is only a small improvement in the local models compared to the full-spectral models, except for amide-N in the *lys3a* material. In order to determine if the relatively small improvement in correlation coefficients and RMSECVs is significant, the models should be evaluated with an independent test set, if a close ranking is desired. Because of the small differences in correlation coefficients and errors, we have decided to further discuss all the intervals having the highest correlation coefficients and the lowest errors selected by si-PLS in Table 3 as a whole in Table 4. Two intervals (numbers 16 and 18) are selected for protein (P) only and one is selected for amide-N (A). Seven intervals are chosen by both types of correlations marked AP. In Table 4, the spectral regions for the intervals are defined and the chemical interpretation given from the literature [15]. We may now also compare to the previously discussed visually selected spectral segment 2270–2360 nm which was indicative for the *lys3a* genotype and which best coincides with the interval number 28 (2290–2358 nm). It is noteworthy that this segment is only in the middle of the ranges of correlations and is not prioritized by the si-PLS selection, although the correlation coefficients are only about 3% lower than those of the optimal segments. Interval number 28 contains mixed information from amino acids (protein) as well as from non-nitrogen components such as cellulose, fat, and C–H bonds. The mixed information is also prevalent for most other selected segments, such as numbers 15, 16, 23, 25, 26, and 27. Only segment 24 seems to be a clear-cut N indicator, while segment 20 indicates carbohydrates and SH groups, and numbers 12 and 18 are C–H indicators.

It can thus be concluded that the NIR spectrum contains repetitive confounded chemical information throughout the spectrum which gives a high degree of redundancy and which in combination with the high precision and repeatability of the measurement explains the versatility and robustness of NIR full-spectrum chemical PLSR predictions brought out by spectroscopic “multimeters” in practice. The consequences of utilizing the multivariate chemical analytical advantage in plant breeding and biotechnology are discussed in Section 4.

4. Discussion

At present, focus in biotechnology tends to be changing from genome sequencing to the concept of the proteome to describe the complement of proteins expressed by a cell tissue, for instance an endosperm. This is called the “the post-genome revolution” in an article in the 16 December 1999 issue of the journal Nature. The complexity of this challenge is lucidly illustrated in Fig. 3a–d comparing the two-dimensional electrophoresis protein patterns from buffer and ethanol extracts from the original Risø 1508 mutant *lys3a* and its isogenic motherline Bomi. They represent about 60% of the total seed proteins with the endosperm as the dominating tissue.

The differences due to just one gene are so great that they cause confusion in comparison of the electropherograms. However, when superimposing the two electrophoresis sets from the albumins/globulins and the prolamins (hordeins), respectively, a pattern of common spots marked with dashed arrows in Fig. 3 proves that the reproducibility is reasonable. The 2 × 19 electropherograms were made for the 15 normal barleys and the four pure mutant lines analyzed for chemical composition in Tables 1 and 2. The differences between the categories, the wildtype and *lys3a*, could be easily discerned in a blind test by visual inspection where a set of anonymized electropherograms were presented in random order. In the literature, the effect of the *lys3a* gene on the synthesis of about 15 proteins has been compared to other relatively well-studied mutants [7,16]. It constitutes only a fraction of the differences demonstrated in Fig. 3a–d. The *lys3a* gene is situated in chromosome 7 [17] and regulates [10] a range of structural genes in other chromosomes, for example in chromosome 5, the loci coding for the hordeins B (Hor 2), C (Hor 1), and D (Hor 3) produced in the endosperm [16]. The first two of those constituting a major part of the hordeins are drastically reduced in *lys3a*, contributing to a total reduction of the hordeins of 85% [18]. At the same time, however, the D-hordeins increase in *lys3a* [18]. The overall reduction of the hordeins was confirmed in the electropherograms when comparing the protein patterns of the alcohol soluble proteins in Fig. 3 (a: Bomi; b: *lys3a*). On the other hand, the protein pattern of the buffer-soluble albumins/globulins in Bomi (Fig. 3c) are in general quantitatively increased in *lys3a* (Fig. 3d) with great qualitative and quantitative differences. Thus, it seems as if the retardation through the *lys3a* gene of the synthesis of the hydrophobic ethanol-soluble hordeins results in a range of hydrophilic buffer-soluble fragments.

With regard to buffer-soluble proteins expressed in the triploid tissue of the endosperm and aleuron in the barley seed, large changes have been confirmed due to the gene *lys3a*. For example, there are decreases in beta-amylase and protein Z [19], while others are increased, e.g. the potential antifungal proteins [20], an amylase/subtilisine inhibitor, a chitinase, and a ribosome-inactivating protein [21,22].

In plant breeding and in biotechnology, there is a great need for screening methods in order to identify genes and gene effects. The genome and proteome concepts introduce a multivariate challenge, where multivariate screening methods such as spectroscopy and data analysis such as chemometrics are central. Recently, Delwiche et al. have demonstrated [6] that it is possible to detect certain wheat–rye chromosomal translocations by NIR spectroscopy. We have described [11] that in a mixed genetic background, it is possible by NIR to distinguish normal barley not only from the drastic high-lysine mutant Risø 1508, but also from the lesser high-lysine mutants Risø 13, 16, 29, and 95 [23] which have not yet been studied from a protein chemistry point of view.

In a classical approach, the biochemist and the biotechnologist tend to focus on specific genes, mechanisms and proteins, while forgetting about the side effects which are considered vital by the geneticist and the plant breeder and are collected in the concept of pleiotropic gene effects. Thus, the Risø mutant 56, in contrast to the regulating gene *lys3a* [24] is a mutation (deletion) in a structural gene Hor 2 in chromosome 5 coding for the B-hordeins, as documented by the absence of the RNA messenger [24]. There are compensatory increases in the C- and the gamma-hordeins [24]. Other pleiotropic effects due to the gene mutant 56, for example on the carbohydrate composition, have not been described but are most likely to occur, because mutants affecting hordein synthesis usually have decreased starch synthesis [7]. If so, the changes in the proteome as well as other changes in other chemical components derived from here could be detected in mutant 56 by NIR spectroscopy, preferably in an isogenic comparison.

Thus, NIR spectroscopy enables a physicochemical fingerprint of the phenotype on the level of phenotypical biological organization which can be compared and analyzed by chemometrics in a PCA, and validated to chemical analyses and knowledge by PLSR. It is thus possible, by defining what is normal barley, to identify and investigate outliers with unknown chemical composition and afterwards define their genetic and chemical status [11]. The material should be grown on the same site, although we have shown in our *lys3a* example that it is possible to separately model the environmental and genetic effects (Figs. 2 and 7). It is clear that a multivariate dataset, collected either by a range of univariate chemical analyses (Tables 1 and 2, Figs. 4 and 5) or more easily by a spectroscopic method (Fig. 2), could facilitate differentiation between genetic and environmental effects compared to a classical approach. The key to utilizing this option is the ability to explore and to model covariance in the datasets by using chemometric methods.

As early as 1930, Bishop [25], in his nitrogen-regulation principle of the Osborne [26] protein fractions of barley, demonstrated a case of covariance implying that as a part of the total protein content, the hordein proteins increased and the albumins and globulins declined when the protein level was increased, for example with nitrogen fertilization. This implied a decline in the total protein content of lysine and essential amino acids and an increase in the content of amides, glutamic acid, and proline due to the composition of these proteins. This mechanism was almost considered a natural law [27] until the discovery of the first high-lysine mutants. Fig. 9A displays the high precision of the negative regression lysine g per 16 gN with total protein content (N × 6.25) for the 15 normal barleys analyzed for amino acids. The introduction of the *lys3a* gene outliers (Fig. 9A) completely changes this picture. It has been confirmed [7] that both the *lys3a* and the *lys1* genes straighten out this correlation, so that now the lysine content of protein with these genes is independent of total protein content. This change is also reflected in comparing the two specific total amino acid patterns as a function of the protein content of the seed, each unique for normal barley and for the high-lysine mutant *lys3a* (Table 1) which is elaborated as a whole in the PCA in Fig. 4.

While the genetic data in our investigation are well defined and hard, the direct and indirect effects of this regulating gene on the endosperm proteome and the phenotype are extremely diverse and multivariate, thus requiring soft mathematical modeling. It is thus in practice impossible to study these effects as a whole without suitable multivariate analytical screening methods like spectroscopy and without chemometric evaluation. We might conclude that gene-dependent, specific, multivariate covariate correlation patterns like those between amino acids as a function of protein content in barley seeds are just as deterministic and environmentally independent a trait as blue and brown eyes in humans.

We will now discuss how the detection of the *lys3a* genotype may work on the NIR spectroscopic level. As seen in Fig. 9B, lysine can be reasonably predicted by full-spectrum NIR (RMSECV O.24 at five PLSR components). A straight 30-interval i-PLS selects interval 28 (RMSECV O.20 at seven PLSR components) which was earlier selected (Fig. 6 and Table 4) as a unique area for visual *lys3a* differentiation from normal barley. A si-PLSR with two synergy elements selects at five PLS components interval 27 together with 28 (RMSECV 0.15 at five PLSR components). However, at seven PLS components, si-PLSR combines the spectral elements 17 and 26 with a minimal error of RMSECV 0.13, approximately half of that for the whole spectral model. These results on lysine, based on 18 spectra, should be confirmed in studies with a larger independent material. However, the purpose of the exercise here is not to differentiate one or two "hot" areas in the NIR spectra from a close ranking list in defining the *lys3a* genotype, but rather to look at several areas with low prediction errors in order to explain how full-spectrum NIR works.

In Fig. 9C, it is seen that lysine mol% is highly negatively correlated to amide-N to N ratio ($r = -0.97$), which points to the possibility that the spectroscopic signature of the amide bond in this material ($n = 125$) could be a good indicator for a low content of lysine. Of the spectral elements, numbers 17, 26, 27, and 28, previously selected by si-PLS correlated with lysine, number 26 (2150–2218 nm) is indicative of amide [15] (Table 4) as well as (unsaturated) fat and the aldehyde group. Area 17 (1520–1588 nm) has information about R–NH_2, starch and the peptide bond, area 27 (2220–2288 nm) about amino acids, cellulose and (unsaturated) fat, while area 28 (2290–2358 nm) includes fiber in addition to the information kept in area 27. In classifying the *lys3a* gene and the wildtype (Figs. 1 and 2) and in predicting chemical constituents such as lysine (Fig. 9B) and protein and amide-N (Fig. 8A–C), the models chosen by the PLSR algorithm do not only rely on direct protein information but also on other different combinations, exploiting the pleiotropic covariate effects of the gene (Table 2). These are due to the influence of the specific proteome on the parts of the cell machinery which are important for the starch, fiber, and fat synthesis (Tables 3 and 4).

It is concluded that the redundancy of biological information on the genotypic DNA sequence level is also represented at the phenotypic level in the dataset read by the NIT/NIR spectroscopic sensor from the chemical/physical fingerprint containing specific

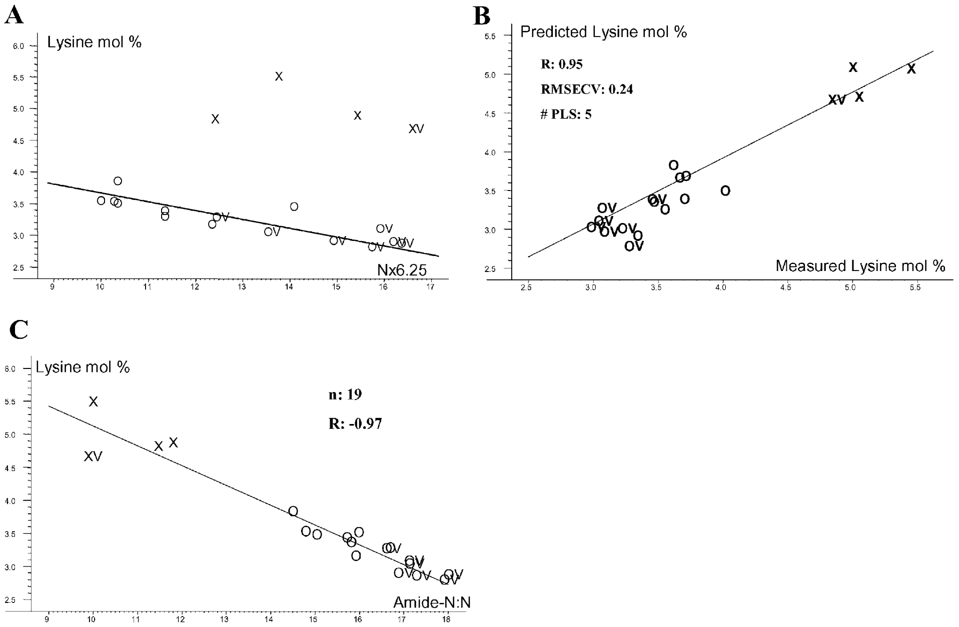


Fig. 9. (A) Data from Table 1 ($n = 19$): the regression line for normal barley (O) for lysine mol% to protein (N × 6.25), X denotes the *lys3a* outliers; (B) PLSR NIR prediction of lysine mol%; one outlier removed ($n = 18$); (C) correlation between lysine mol% and the ratio amide-N to total N.

genetic information which can be encoded by chemometrics. NIR spectroscopy may in the future find an extended use in selecting transformants not only coding for genetically engineered proteins [28], but also selecting for other genes by exploiting the pleiotropic effects as markers instead of using antibiotic-resistant genes. This could be done with great sensitivity and precision in an isogenic background with material grown under the same environmental conditions.

In barley, the *lys3a* gene exerts a negative pleiotropic effect on the starch content. Conventional breeding by changing the gene background has improved starch quantity without loss in protein quality [29]. By using NIR spectroscopy evaluated by chemometric methods, pleiotropic effects can be quantified as a whole and explained by chemical validation. For example, NIR spectroscopy combined with chemical validation makes it possible to define a pleiotropic covariate complex in high-lysine barley breeding for the *lys3a* gene. Thus, in a high-lysine barley breeding program, a chemometric model of selection in cross-breeding populations by NIR could be defined which neutralizes the negative parts of the pleiotropic complex by restructuring the multigene background. Classical, exploratory plant breeding could thus be upgraded to a high-tech analytical status and the cooperation with normative biotechnology improved.

Our example is a special case of proteome dynamics cast as a covariate chemical imprint in the desiccated seed endosperm tissue. The spectroscopic and chemometric screening concept presented here should have great advantages in isolating mutants and gene transformants by revealing covariate gene indicators in studying the dynamics of growth in cultures of cells and microorganisms.

The above-cited post-genome revolution by the proteome claimed by the journal Nature in December 1999 will not constitute the end point of biological science. Already now, we are envisaging the possibility of utilizing the high precision and multivariate advantage of spectroscopy to separately model the covariate expressions of genetic and environmental variation by chemometrics. This enables us to overview and non-destructively, separately model the genetic and environmental variation at the highest level of biological organization the chemistry of the intact phenotype, the chemotype, within the limits of our sensors and our chemometric evaluation methods.

Acknowledgements

The research was kindly financed by the Danish Cereal Network, Department of Food and Agriculture. The authors thank Anna Marie Hansen, Peter Henriksen, and Kirsten Wilms for the laboratory work and Gilda Kischinovsky for assistance with the manuscript. The barley seed material emerges from a 40-year research cooperation with colleagues at the Swedish Seed Association, Svaløf, Sweden, and with The Risø National Laboratory and The Carlsberg Research Laboratory in Denmark. Preben Holm at the Danish Institute of Agricultural Sciences, Flakkebjerg, and Gary Fulcher from the University of Minnesota, St. Paul, USA, have contributed inspiring discussions.

References

[1] W. Herschel, Phil. Trans. R. Soc. (Lond.) 284 (1800) 284–292.
[2] W.W. Abney, R.E. Festing, Phil. Trans. R. Soc. (Lond.) 31 (1881) 416.
[3] P.C. Williams, K.H. Norris, Near-Infrared Technology in Agriculture and Food Industries, American Association of Cereal Chemists, St. Paul, MN, USA, 1987.
[4] K.H. Norris, Agric. Eng. (St. Joseph, MI, USA) 45 (1964) 370.
[5] H. Martens, S.Aa. Jensen, in: J. Holas, J. Kratochvil (Eds.), Proceedings of the 7th World Cereal and Bread Congress, Prague, June 1982, Elsevier, Amsterdam, 1983, p. 607.
[6] S.R. Delwiche, R.A. Greybosch, C.J. Peterson, Cereal Chem. 76 (1999) 255.
[7] L. Munck, in: P.R. Shewry (Ed.), Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology, CAB International, Wallingford, UK, 1992, p. 573.
[8] L. Munck, K.E. Karlsson, A. Hagberg, B.O. Eggum, Science 168 (1970) 985.
[9] J. Ingversen, B. Køie, H. Doll, Experientia 29 (1973) 1151.
[10] M. Blom Sørensen, M. Muller, J. Skeritt, D. Simpson, Mol. Gen. Genet. 250 (1996) 750.
[11] L. Munck, in: R. Von Bothmer, et al. (Eds.), Diversity in Barley, Elsevier, Amsterdam, 2001, Chapter XII, manuscript accepted, in preparation.
[12] V. Barkholt, A.L. Jensen, Anal. Biochem. 177 (1989) 318.
[13] S. Jacobsen, L. Nesic, M. Petersen, I. Søndergaard, Electrophoresis 22 (2001) 1242–1245.
[14] L. Nørgaard, A. Saudland, J. Wagner, J. Pram Nielsen, L. Munck, S.B. Engelsen, Appl. Spectrosc. 54 (2000) 413.
[15] B.G. Osborne, T. Faern, P.H. Hindle, Practical NIR Spectroscopy, 2nd Edition, Longman, Harlow, UK, 1993.
[16] M. Kreis, P.R. Shewry, in: P.R. Shewry (Ed.), Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology, CAB International, Wallingford, UK, 1992, p. 319.
[17] K.E. Karlsson, Barley Genet. Newslett. 7 (1977) 40.
[18] M. Kreis, P.R. Shewry, B.G. Forde, B.G. Miflin, Biochem. Gen. 22 (1984) 231.
[19] J. Hejgaard, S. Boisen, Hereditas 93 (1980) 311.
[20] R. Leah, H. Tommerup, I. Svendsen, J. Mundy, J. Biol. Chem. 266 (3) (1990) 1564.
[21] J. Mundy, J. Hejgaard, A. Hansen, L. Hallgren, K.G. Jørgensen, L. Munck, Plant Physiol. 81 (1986) 630.
[22] R. Leah, Barley Seed Antimicrobial Proteins, Dissertation, Department of Microbiology, University of Copenhagen, 1991.
[23] H. Doll, in: W. Gottschalk, H.P. Muller (Eds.), Seed Proteins: Biochemistry, Genetics, Nutritive Value, Martin Nijhoff, The Hague, 1985, p. 205.
[24] M. Kreis, P.R. Shewry, B.G. Forde, S. Rahman, B.J. Miflin, Cell 34 (1983) 161.
[25] L.R. Bishop, J. Inst. Brew. 36 (1930) 352.
[26] T.B. Osborne, The Vegetable Proteins, Longmans & Green, London, 1920.
[27] W. Postel, Zuchter 26 (1956) 211.
[28] M. Torrent, I. Alvarez, M.I. Geli, I. Dalcol, D. Ludevid, Plant Mol. Biol. 34 (1997) 139.
[29] K. Bang-Olsen, B. Stilling, L. Munck, Barley Genetics V, Proceedings of the Fifth International Barley Genetics Symposium, University of Okayama, Japan, 1987, p. 865.

Paper 6

Development of non-destructive screening methods for single kernel characterisation of wheat

Jesper Pram Nielsen, Dorthe Kjær Pedersen and Lars Munck

Abstract

The development of non-destructive screening methods for single seed protein, vitreousness, density and hardness index has been studied for single kernels of European wheat. A single kernel procedure was applied involving, image analysis, Near Infrared Transmittance (NIT) spectroscopy, laboratory density determination, Single Kernel Characterization System (SKCS) and finally Kjeldahl protein determination on the crushed single kernels.

Single kernel NIT spectroscopy showed excellent ability to determine protein content, and some ability for determination of single kernel vitreousness. Non-destructive determination of single kernel density, either based on NIT spectroscopy or based on image analysis and kernel weight, needs to be further improved for practical use.

The use of SKCS hardness index as a true single kernel hardness reference in a NIT prediction model resulted in a poor predictability. However, by applying an averaging approach, in which single seed replicate measurements are mathematically simulated, a very good NIT prediction model was achieved. This suggests that the single seed NIT spectra contain hardness information, but that a single seed hardness method with higher accuracy is needed, in order to achieve a good NIT prediction model for single kernel hardness.

Introduction

The purpose of this paper is to apply a combinatory single seed approach involving several types of single seed measurements on the same individual seeds for an improved wheat characterisation, with special emphasis on single kernel protein, vitreousness, density and hardness.

Protein content, kernel density in terms of test weight, and kernel vitreousness by visual inspection are normally used in the miller's quality evaluation of wheat for milling. Protein content largely determines the end use quality, and premiums are often offered on high protein wheat. Test weight reflects kernel size and density, and should be above a certain level in order to secure a good flour yield. The vitreousness is used for evaluation of millability, even though the relationship between vitreousness and hardness is not straightforward. Vitreousness and/or hardness affects the milling processing of wheat, including tempering of the grains, flour yield and the end-use properties such as particle size distributions and the amount of damaged starch. Grain hardness is mainly determined by the degree of adhesion between the starch granules and the protein matrix, with a tight adhesion of the starch granules in the hard wheat and a weaker adhesion in soft wheat. Even though wheat can be divided into genetically soft and hard, a substantial variation in texture is seen within the two classes, and the apparent vitreousness of the wheat is therefore used by the millers in their evaluation of millability.

Wheat quality evaluation has traditionally been performed on bulk samples, which implies that the characteristics of the individual kernels within the sample is lost, and thereby the opportunity to evaluate sample homogeneity. In seed sorting and grading by size, form and density for better and more uniform quality, the single seed is the functional unit to be investigated. New developments in instrumentation have made single kernel characterisation possible, and for some quality parameters rapid enough, to become a valuable tool for homogeneity evaluation in the cereal industry. The Single Kernel Characterization System (SKCS) 4100 (Perten Instruments Inc., Reno, NV, USA) is an example of such an instrument for rapid, albeit destructive, measurement of single kernel hardness, weight, diameter and moisture content (Martin et al., 1993). The single kernel measurements are normally conducted on 300 single kernels in a bulk sample in order to classify the sample into soft, hard or mixed wheat.

One of the limitations of destructive single seed analysis is that several readings on the same kernels are impossible. It therefore becomes difficult to differentiate between instrument variability and kernel-to-kernel variability. By using non-destructive single seed analyses these problems could be circumvented. Additionally, fast and non-destructive single kernel quality analyses would be valuable tools in plant breeding for quality selection in early generations and for single kernel quality evaluation within the heads.

Near infrared spectroscopy on single kernels fulfils these requirements and the technique has been used for several single kernel applications. Near Infrared Transmittance (NIT) spectroscopy has been reported for determination of oil in maize (Orman and Schumann, 1992) and meadowfoam (Patrick and Jolliff, 1997), protein in wheat (Delwiche, 1995) and soybeans (Abe et al., 2000) and for wheat hardness (Delwiche, 1993). Near infrared reflectance spectroscopy has similarly been applied for wheat classification (Delwiche and Massie, 1996), for determination of single seed protein (Delwiche, 1998; Delwiche and Hruschka, 2000), for differentiation between vitreous and non-vitreous durum wheat kernels (Dowell, 2000) and for assessment of heat-damaged wheat kernels (Wang et al., 2001).

Image analysis is another method for fast non-destructive characterisation of kernels. Image analysis has been used for discrimination between kernels of different species (Chtioui et al., 1996), discrimination between wheat classes and varieties (Zayas et al., 1986) and, used in combination with physical measurements, for variety identification (Zayas et al., 1996). Berman et al. (1996) used the method for screening of flour milling yield in wheat breeding.

This investigation involves a combination of image analysis; NIT spectroscopy, hardness analysis (SKCS), protein analysis as well as a simple laboratory density analysis applied on single kernels of European wheats. The paper includes a survey of the use of non-destructive screening methods for prediction of single kernel protein, vitreousness, density and hardness.

Material and Methods

Samples:

Bulk samples of 43 different wheat cultivars or mixtures of cultivars in common use, from two different locations in Denmark (Jutland and Funen) were collected, representing both genetically hard and soft varieties. In order to select full developed kernels, the samples were screened on a 2.2 mm screen and the fractions above 2.2 mm were stored separately in plastic bags. Five kernels were chosen randomly from each of the 86 bulk samples to make up the calibration set (430 kernels in total). Another ten kernels from each of 11 of the 86 bulk samples (11 cultivars from Funen) were selected as the test set (110 kernels in total). Since the measurements of the kernels in the calibration set showed no significant differences between the two locations, it was chosen only to use test set kernels from one of the locations.

Single kernel measurements:

The single kernels were put through the following sequence of measuring steps. The kernels were analysed one by one with their identity retained during the measurement procedure.

GrainCheck:

Grain morphology was measured by digital image analysis using a GrainCheck^{TM} 310 instrument (FossTecator, Höganäs, Sweden). The instrument was used for single kernel characterisation by manually placing each kernel under the RGB camera from which the kernels were imaged and from which several morphological and color characteristics were automatically assessed. In this investigation the following nine kernel characteristics were registered from the instrument and used in the data analysis: kernel width, kernel length, roundness, area, volume, red reflectance, green reflectance, blue reflectance, and total light reflectance.

NIT Spectra:

After the GrainCheck analysis, the single kernels were moved to an Infratec 1255 Food and Feed Analyzer (FossTecator, Höganäs, Sweden). Each kernel was placed in a single seed sample cassette with slots for 23 single kernels, and near infrared transmittance (NIT) spectra in the range 850-1050 nm were automatically recorded. Spectra were recorded three times on each kernel and the average of the three spectra was used. The position of the kernels in the sample cassette was manually changed between each of the three measurements. The time required for scanning (single scan) 23 single kernels in the cassette was about 90 s.

Single kernel density:

A laboratory single kernel density measurement was developed and applied to the 110 test set kernels prior to the SKCS analysis. The kernels were individually weighed to the nearest 0.1 mg using a Mettler/Toledo scale (Type AB204). When immersing a wheat kernel in water, the weight of the displaced water divided by the density of the water equals the kernel volume. This measurement was carried out by using the equipment shown in Figure 1, which was specially designed for the purpose. A beaker containing water at 20°C was placed on the Mettler/Toledo scale.

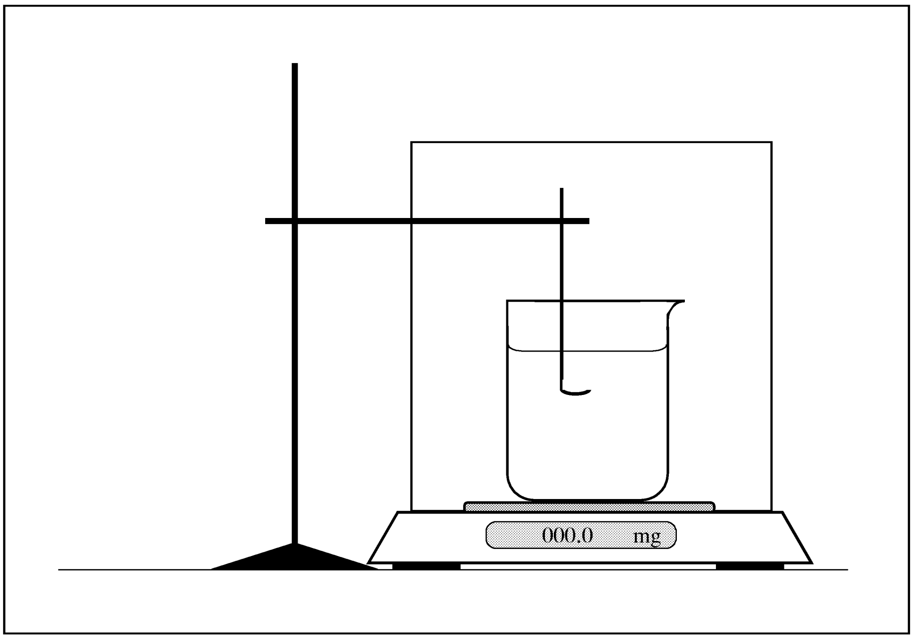


Figure 1. Illustration of the method for determination of single kernel volume.

A single kernel holder (modified sample spoon) was mounted on a rack outside the scale chamber (without touching the scale) with the kernel holder end immersed in the water. The scale was tared and the kernel (one at the time) was placed in the holder using a needle. The weight of the water displaced by the volume of the kernel was recorded immediately after, in order to avoid too much water uptake by the kernel. After the analysis, the kernels were dried for 16 hours at 30°C, and checked to have returned to the same weight as prior to the volume measurements.

Having determined the kernel volume from the weight of the displaced water, the single kernel density (in g/cm^3) is subsequently calculated by dividing the kernel weight (g) by the volume (cm^3). Prior to the single seed analyses the volume method was tested on 10 glass beads differing slightly in volume. The average deviation between the "real" volume and the volume determined using the method shown here was 0.0004 cm^3 for an average of 0.0142 cm^3, i.e. an error of 2.8 %.

Perten SKCS analysis:

The kernels were subsequently analysed using a Single Kernel Characterization System (SKCS) 4100 (Perten Instruments Inc., Reno, NV, USA). The SKCS measures a single kernel hardness index (HI), moisture content (%), diameter (mm) and weight (mg). Normally, the SKCS analysis is carried out on a small bulk sample (300 kernels), but in this experiment the single kernels were fed one by one into the vacuum wheel in order to retain their identity. The normal container for collecting the crushed kernels was removed, and the single kernel grist from the individual kernels was collected in a small container and used without further grinding for determination of single seed protein according to Kjeldahl.

Protein determination on single kernels:

Single kernel nitrogen content was finally determined directly by a modified Kjeldahl (1883) method according to the AACC Method 46-12. The protein content is reported as percent in dry matter calculated using the moisture content measured by the SKCS instrument. Prior to the single kernel analysis, the method was tested on samples of 30-40 mg wheat flour. The analytical error in terms of standard deviation of 20 replications amounted to 0.16 % (percent protein content in dry matter).

GrainCheck data, NIT spectra, SKCS data and protein content were then recorded for each kernel, and single kernel density was determined on each of the kernels in the test set. A disadvantage of destructive single seed analysis is that if a measurement fails, there is no sample left for a second analysis. Here, a few of the SKCS, protein and volume analyses failed, and the following results and discussion are therefore based on a slightly reduced number of kernels. The calibration set consists of 415 out of the original 430 kernels, while the test set of 110 kernels gave valid data for 108 kernels, except for the density measurements where valid results were obtained for only 99 kernels.

The mean and range of all the 14 non-spectral single kernel characteristics for the calibration set kernels and the test set kernels are given in Table I.

Table I: Mean and range of the recorded single kernel characteristics.

Method	Parameter	Calibration set (n=415)			Test set (n=108)			Total (n=523)		
GrainCheck		Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
	Width (mm)	3,7	2,3	5,0	3,8	2,5	4,7	3,7	2,3	5,0
	Length (mm)	6,2	5,0	7,5	6,1	5,0	7,2	6,2	5,0	7,5
	Roundness (AU)[a]	0,50	0,25	0,83	0,54	0,31	0,82	0,51	0,25	0,83
	Area (mm^2)	16,9	9,4	25,8	17,1	10,2	24,8	17,0	9,4	25,8
	Volumen (mm^3)	40,8	14,6	82,8	42,6	16,6	76,0	41,2	14,6	82,8
	Red	46,4	31,9	62,4	44,0	25,9	60,5	45,9	25,9	62,4
	Green	33,6	22,7	46,5	31,8	17,6	43,9	33,3	17,6	46,5
	Blue	24,4	17,4	34,0	23,2	14,8	31,1	24,1	14,8	34,0
	Intensity	34,8	24,2	47,4	33,0	19,5	44,9	34,4	19,5	47,4
SKCS										
	Weight (mg)	45,1	24,5	68,0	45,1	24,1	69,3	45,1	24,1	69,3
	Diameter (mm)	2,9	1,7	4,6	3,0	1,7	4,3	2,9	1,7	4,6
	Moisture (%)	11,9	10,4	13,3	11,0	10,0	11,6	11,7	10,0	13,3
	Hardness (HI)	44,0	-21,4	101,5	32,3	-28,8	82,2	41,6	-28,8	101,5
Reference										
	Protein (% DM)	10,0	6,8	15,2	9,8	7,0	17,0	10,0	6,8	17,0
	Density (g/cm^3)				1,16	0,99	1,25			

[a] Values in the range of 0-1. A perfect circle has roundness=1, while a very narrow elongated object has roundness close to 0.

Data analysis:

Partial Least Squares Regressions (PLSR) (Martens and Næs, 1993) were performed using Unscrambler version 7.6 (CAMO A/S, Norway) in order to predict a given quality parameter (**y**) from fast acquirable **X** data. The multivariate prediction results are presented and discussed as correlation coefficients (r) between predicted and measured values, and prediction error in terms of Root Mean Square Error of Prediction (RMSEP) for true test set predictions, and Root Mean Square Error of Cross Validation (RMSECV) for cross-validated results. Relative predictions errors (RE) reported in percent are calculated by dividing the prediction errors (RMSECV or RMSEP) by the range (max. - min. value) of the given parameter.

Results and discussion

Single kernel protein:

The statistics of the Kjeldahl protein determination are listed in Table I. The single seed protein content ranges from 6.8% to 17.0% for all the analysed kernels, and thus in principle covers the whole range of end-use requirements from low-protein wheat for crackers to high-protein wheat for bread making. In order to evaluate and utilise this single seed protein variation, a spectroscopic method would be useful. For this purpose, we use single seed NIT spectra recorded on each of the 523 wheat kernels in the spectral region 850-1050 nm. The NIT spectra of the 523 single wheat kernels are shown in Figure 2 as both raw spectra (a) and scatter-corrected spectra (b), applying a combination of second derivative followed by Multiplicative Scatter Correction (MSC) (Geladi et al., 1985).

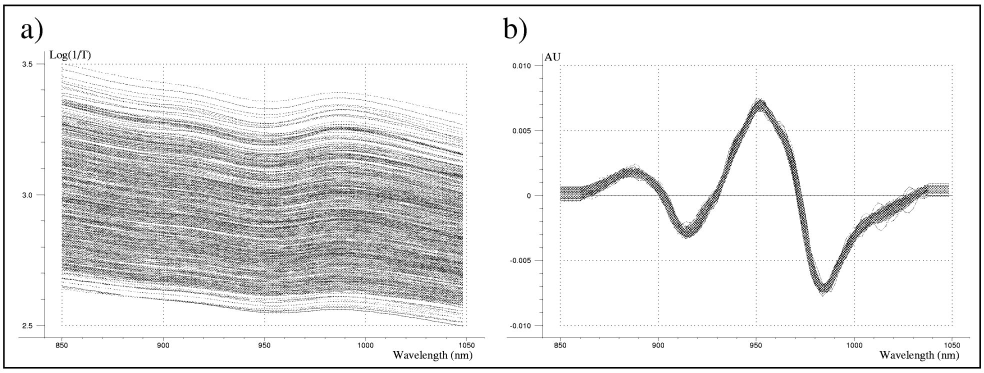


Figure 2. Single seed NIT spectra of 523 wheat kernels shown as (a) raw spectra and (b) corrected spectra using second derivative followed by MSC.

This combined scatter correction has been discussed by de Noord (1994) and applied to single seed NIT spectra by Delwiche (1995). The raw spectra show large intensity offsets, as well as less clear multiplicative effects. These scatter effects are probably due to differences in kernel size and texture together with kernel orientation in the single seed cassette. With respect to the scatter-corrected spectra (Figure 2b), it is evident that the spectral scatter has been corrected for, and thereby more spectral emphasis could be focused to represent chemical composition, e.g. the level of water, starch and protein content in the kernels. Delwiche (1995) showed that the combination of second derivative of the single seed NIT spectra followed by MSC gave the best predictions. Our results are in agreement with this finding because raw spectra, first derivative spectra, second derivative, MSC or MSC followed by second derivative corrected spectra (data not shown) were less efficient in a prediction model. The issue of scatter in single seed NIT spectra, including suggestions for more general and powerful pre-transformations, is further investigated by Pedersen et al. (2002).

A prediction model for protein content was developed based on single seed NIT spectra corrected by the second derivative followed by MSC. The cross-validated calibration model using 5 PLSR components including 415 single kernel spectra is shown in Figure 3a. This calibration model is used for independent prediction of the 108 test set kernels (Figure 3b). The relatively low number of PLSR components (5) as compared to other PLSR models in the near infrared range implies a simple and thus robust model. The prediction error (RMSEP) of 0.48 % protein when tested independently on 108 new kernels also indicates a good and robust calibration model. Our results for single seed protein determination are comparable to results reported earlier using near infrared transmittance (850-1050 nm) (Delwiche, 1995) and near infrared reflectance (1100-2498 nm) (Delwiche, 1998).

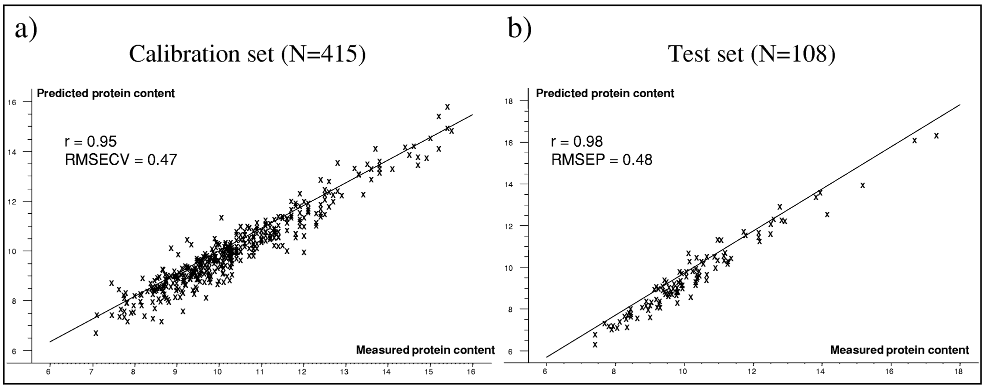


Figure 3. Predicted versus measured plot of a 5 PLSR component regression model for single seed protein using scatter-corrected NIT spectra for (a) the calibration set and (b) the subsequent prediction of the test set kernels.

Single kernel vitreousness:

Kernel vitreousness is normally determined by visual inspection, where vitreous kernels appear glassy and translucent whereas non-vitreous kernels appear starchy and opaque. Vitreouness is mainly controlled by nitrogen availability in the field as well as temperature during grain filling (Pomeranz and Williams, 1990). Vitreous kernels are often harder and have higher protein content. In this investigation we apply RGB image analysis by the GrainCheck instrument in order to provide a fast and objective analysis of vitreousness. As a pre-test to the current investigation we analysed vitreous and non-vitreous kernels (selected by visual inspection) on the image analyser (GrainCheck). Among the registered color data it was found that especially the red color reflectance differentiated well between vitreous and non-vitreous kernels. The red reflectance from GrainCheck was therefore selected as a quantitative measurement of vitreousness and denoted "GrainCheck vitreousness". The more vitreous the kernel, the lower the red reflectance and vice versa, i.e. the higher the number, the more non-vitreous the kernel appears. A single seed correlation coefficient of -0.63 (Table II) between protein content and GrainCheck vitreousness shows that the kernels with high protein kernels tend to be more vitreous.

A PLSR model (not shown) was computed using the raw NIT spectra for the prediction of GrainCheck vitreousness in order to see if the NIT spectra contained information regarding the GrainCheck vitreousness. The correlation coefficient between measured and predicted GrainCheck vitreousness was 0.76 with a prediction error of 4.5 AU. A subsequent test of this model on the 108 test kernels confirmed the calibration results (r=0.76, RMSEP=4.6 AU). Even though the NIT model is based on 6 PLSR components, most of the spectral NIT information is simply based on the level of absorbance. This can be concluded, since the first score from a Principal Component Analysis (not shown) on the raw NIT spectra (Figure 2), mainly representing differences in optical densities (offset) correlates well (r=0.71) with GrainCheck vitreousness. The raw NIT spectra thus contain information regarding the GrainCheck vitreousness.

Single kernel density:

Kernel density is an important parameter in the milling industry, which is normally determined on bulk samples as test weight. The test weight measurement is greatly influenced by kernel packing, kernel size and kernel density, without differentiation between those factors.

Table II. Correlations coefficients (r) between protein content, density, GrainCheck vitreousness and SKCS hardness.

	Correlation coefficient (r)
Protein content vs. GrainCheck vitreousness[a]	(-) 0.63
Protein content vs. density[b]	0.65
Protein content vs. SKCS hardness[a]	0.38
SKCS hardness vs. GrainCheck vitreousness[a]	(-) 0.55
SKCS hardness vs. density[b]	0.34
Vitreousness vs. density[b]	(-) 0.53

[a]: N = 523 kernels
[b]: N = 99 kernels

Utilising differences in kernel density by grading for a better and more uniform quality on for example gravity tables, the link between single kernel density and other single kernel quality parameters is essential, in order to predict if a given sample is worthwhile sorting for density. For instance, there should be a link between single kernel density and single kernel protein in order to be able to sort for higher protein content by indirectly sorting for density.

The single kernel density in the test set of 99 kernels ranges from 0.99 g/cm^3 to 1.25 g/cm^3. In this material of European wheats, a correlation coefficient of 0.65 (Table II) between protein content and density and a correlation coefficient of -0.53 between GrainCheck vitreousness and density was seen (Table II). A single seed correlation coefficient of 0.65 between protein and density would probably be too low to be able to sort for protein by use of density grading on a gravity table.

The "Archimedes" procedure developed and used for single seed volume analysis in this investigation is rather tedious and it was of interest to investigate whether the much more rapidly acquirable NIT or GrainCheck data could be used for good volume and density determinations. The GrainCheck provides a calculated value of kernel volume based on a 2D-image. Densities derived from these calculated volumes gave, however, a poor correlation (r=0.07) to the real densities based on "Archimedes". This low correlation is most likely due to the approximation of a 3D-volume based on a 2D-image, which even if it gives a correlation coefficient of 0.9 to the "real" volume (Archimedes) is not sufficiently accurate to provide the basis for an accurate measurement of single kernel density.

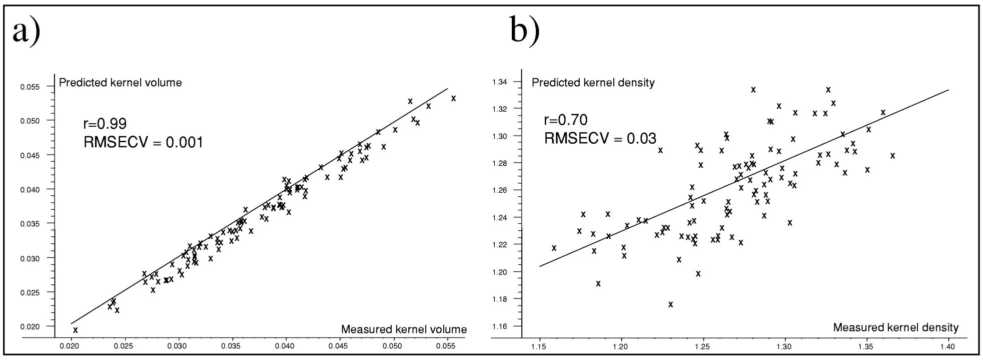


Figure 4. a) Predicted versus measured plot of a PLSR model for kernel volume using the nine GrainCheck variables plus single kernel weight. b) Predicted versus measured plot of a PLSR model for kernel density using the nine GrainCheck variables plus single kernel weight.

A second approach, in which the nine GrainCheck variables (see Table I) plus the kernel weight were used as **X** in a PLSR model, gave a good prediction of the single kernel volume (Figure 4a). This combination of image analysis data with kernel weight gives an excellent, rapidly acquirable estimate of the single kernel volume (r=0.99, RMSECV=0.001cm^3) using full cross-validation (N=99). The subsequent calculation of the single kernel density based on this predicted volume provides a considerably better estimate of kernel density, but still only a correlation coefficient of 0.68, as compared to the 0.07 above, with a prediction error of 0.04 g/cm^3 (plot not shown).

Thirdly, by directly using the nine GrainCheck variables plus the kernel weight for PLSR prediction of density, the results can be improved slightly, giving r=0.70 and a lower prediction error (RMSECV=0.03 g/cm^3) (Figure 4b).

In a final approach, it was investigated whether the NIT spectra contained information, which could be used for prediction of single kernel density. For a PLSR model using the raw NIT spectra for the prediction of the kernel density, the correlation between measured and predicted density gave 0.63 with a cross-validated prediction error of 0.035 g/cm^3. An attempt to combine GrainCheck and NIT data for an improved prediction of kernel density was not successful.

Single kernel hardness:

We have now provided data on the single kernel basis for protein content, kernel density and apparent vitreousness, the tools normally used by the miller for wheat quality evaluation. Hardness is also used for classification of wheats and its quality in relation to different end uses. It was of interest to investigate to what extent hardness added any further information to the structural characterisation of wheat in addition to kernel vitreousness and density.

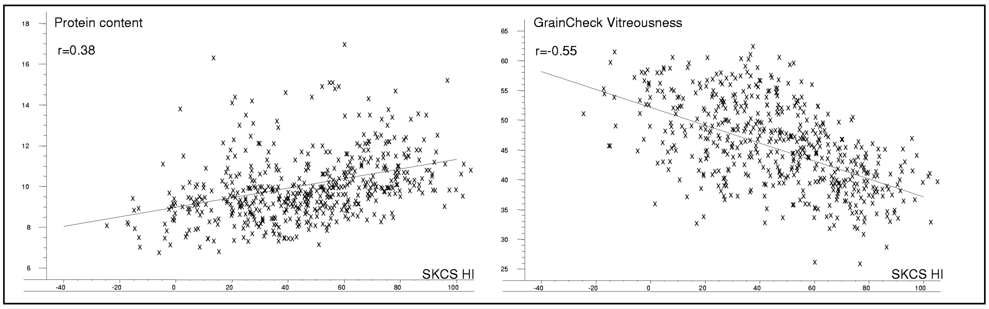


Figure 5. Scatter plots of (a) single seed (N=523) SKCS hardness versus protein content and (b) single seed SKCS hardness versus GrainCheck vitreousness.

In this investigation each kernel was fed separately into the SKCS in order to retain its identity and thereby explore the link between SKCS HI and other single kernel quality parameters. The range in SKCS HI for the analysed kernels is shown in Table I. Figure 5 shows a scatter plot of single seed SKCS HI versus a) the protein content and b) the GrainCheck vitreousness. A low correlation (r=0.38) between protein content and SKCS HI indicates that the SKCS HI is nearly independent of the kernel protein content in this wheat material. This is surprising, as it is often assumed that high protein wheat kernels tend to be harder. The low correlation between single kernel Kjeldahl protein content and SKCS hardness might be explained by the fact that the kernels originate from a range of genotypes, and that the link between seed protein and seed hardness is seen in some genotypes but not in others. The low number of kernels (10) within each variety in this experiment, however, does not allow for investigation of the correlations within each of the varieties.

A higher, yet still low, correlation (r=-0.55) is seen between the GrainCheck vitreousness and the SKCS HI (Figure 5b). Table II summarises the correlations between protein content, density, GrainCheck vitreousness and SKCS HI. Only a small portion of the SKCS HI information seems to be explained in protein content, vitreousness or density as seen by the relatively low correlations.

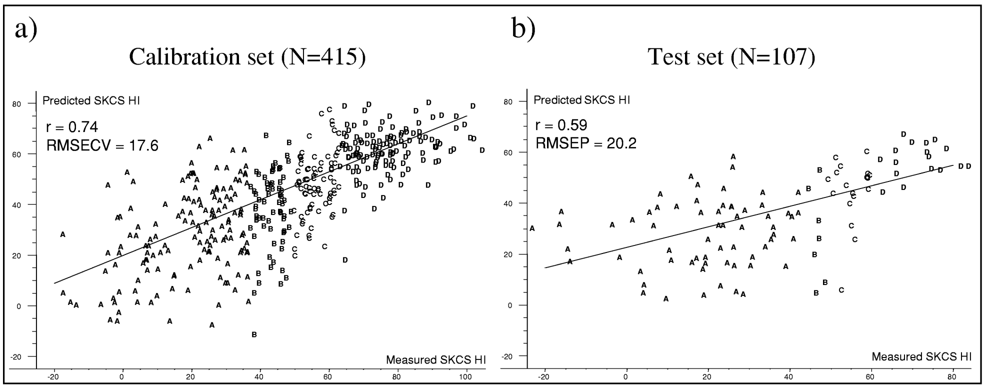


Figure 6. Predicted versus measured plot of a 6 PLSR component regression model for single seed SKCS hardness using (a) the raw NIT spectra for the calibration set and (b) the subsequent prediction of the test set kernels

In bulk, NIT has been successfully applied for prediction of texture in wheat. Williams (1991) concluded that a bulk NIT measurement was capable of predicting whole-wheat kernel texture with precision equal to that of the Particle Size Index

(PSI) method and slightly better than the NIR method. Delwiche (1993) reported on the use of single kernel NIT measurements for hardness determination. When calibrating single seed NIT spectra against bulk hardness data, he found that NIT spectra of single seeds had some ability to determine wheat hardness.

Here we attempt to develop a PLSR model between single seed NIT spectra and true single seed hardness data, namely the SKCS hardness index. In general, we achieve better prediction models for kernel hardness using the raw NIT spectra compared to scatter corrected spectra, which agrees with the findings of Delwiche (1993). A prediction model (6 PLSR components) for SKCS HI based on the raw single seed NIT spectra using segmented cross validation was performed on the calibration kernels. A reasonable calibration is achieved (r=0.74, RMSECV=17.6 HI) as shown in Figure 6a. This calibration was subsequently used for HI prediction of 107 of the original 108 test set kernels (Figure 6b). A low correlation coefficient of 0.59 and a high prediction error of 20.2 HI was achieved. This prediction error corresponds to 20% of the hardness range and thus limits the practical use. In Figure 6a and Figure 6b the samples are labelled according to the hardness groups, where "A" is soft (HI<33), "B" is semi-soft (33<HI<46), "C" is semi-hard (46<HI<59) and "D" is hard (HI>59). It is apparent that the soft kernels (denoted “A”) give a more scattered picture in the plots, which means that the hardness index of these kernels are more difficult to predict. However, an exclusion of the soft “A” kernels did not improve the results.

Various aspects have been considered when interpreting the reason for the relatively poor NIT prediction of SKCS HI we achieve in this investigation. First, there might not be a link between single seed NIT spectra and single seed kernel hardness, but, as mentioned above, earlier reports have demonstrated the use of NIT spectroscopy on whole-wheat kernels for hardness determination. Secondly, irrelevant noise in the NIT spectra (**X**) and the SKCS hardness data (**y**) might impair the model. Our single seed NIT spectra are averages of three spectra recorded on each kernel. As shown earlier, these spectra correlate very well with kernel protein, so the quality of the NIT spectra seems to be satisfactory. On the other hand, the single seed HI, as determined by the SKCS, might be too inaccurate and thereby problematic as y-values in a NIT prediction model. Since the SKCS HI measurement is destructive, multiple HI readings on the same kernel are not possible and an average of replicate readings is thereby impossible to obtain. This essential condition also makes it difficult to quantify the uncertainty of the instrument measurement. If it was possible to prepare a uniform set of kernel shaped particles from a polymer material, it could be an opportunity to estimate the single kernel uncertainty of the SKCS HI measurements. In the current investigation, a possible way to investigate this problem of uncertainty is to mathematically simulate replicate measurements by averaging across single kernels that are nearly identical. First, we have applied such an averaging approach for the NIT model to protein content where we are certain of both the NIT spectra and the Kjeldahl protein content determinations. Since this method requires a great number of samples, we use all the 523 analysed kernels. The NIT spectra and corresponding protein content values are sorted according to protein content. As a start, a PLSR model is developed on the basis of all the 523 calibration kernels. Then, the sorted data are averaged across two kernels. Since the kernel data are sorted according to protein content, the two-kernel average is an average, which might be taken as an average of two duplicated analyses on one kernel. A subsequent PLSR model is then developed for the 262 averaged data objects (averaged kernels). This procedure is repeated another 4 times in which PLSR models are developed averaging across 1 (N=523), 2 (N=262), 4 (N=131), 8 (N=66), 16 (N=33) and 32 (N=17) kernels, respectively. For each model the percent of non-explained variation of the total variation is calculated. The trend of non-explained variation of the protein data for the different PLSR models can then be evaluated (Figure 7, dotted line). In an ideal situation i.e. with no noise in the NIT spectra and with determinations of Kjeldahl protein content without any errors, together with a perfect description of the protein content by the NIT spectra, a horizontal line at an ordinate value of 0 would have appeared. In a situation in which we only have model error, i.e. not perfect description of the protein content by the NIT spectra, but still with no noise in the NIT spectra and Kjeldahl protein content measurements, we would expect a horizontal line at a certain level above an ordinate value of 0. The decrease in non-explained variation when averaging (moving from left to right in the plot) represents the noise and errors in the NIT spectra and in the Kjeldahl protein content determinations, reaching a horizontal level representing only model error as mentioned above. As seen from the dotted line, approximately 13% of the protein data variation is not explained by the NIT PLSR model using all kernel data (original single kernel data), but already after averaging over 4 kernels (2^2), a nearly horizontal line is appearing at approximately 4% non-explained variation. This means that after 4 simulated replicates, nearly all data noise and errors have been eliminated.

The exact same strategy was applied to the NIT model for SKCS HI (only 522 out of the original 523 kernels had valid data and were used). The results are shown in Figure 7 (solid line). It is evident that the non-explained variation in the HI model is considerably higher than for the protein model. As much as 50% of the HI data variation is not explained by the NIT PLSR model using all kernel data, and even after averaging 32 kernels (2^5) the curve is still declining slightly, reaching a level around 15% non-explained variation. When comparing the two models which are based on the exact same NIT spectra, it is apparent that the decrease in non-explained variation when averaging is much smaller for the protein model compared to the HI model, thus indicating considerably higher measurement errors in the HI measurement. Table III summarises the averaging approach in terms of correlation coefficients (r) and RMSECV for the protein content and SKCS HI prediction models. It is seen that by averaging 32 times a good prediction model for HI is developed reaching a correlation coefficient of 0.93 and a prediction error of 10.4 HI, which corresponds to 10% of the range. This good model suggests that the raw NIT spectra can be used for single seed prediction of SKCS HI. However, the results also show that the single SKCS HI values are not sufficiently accurate to be used as reference values in a NIT-based prediction model.

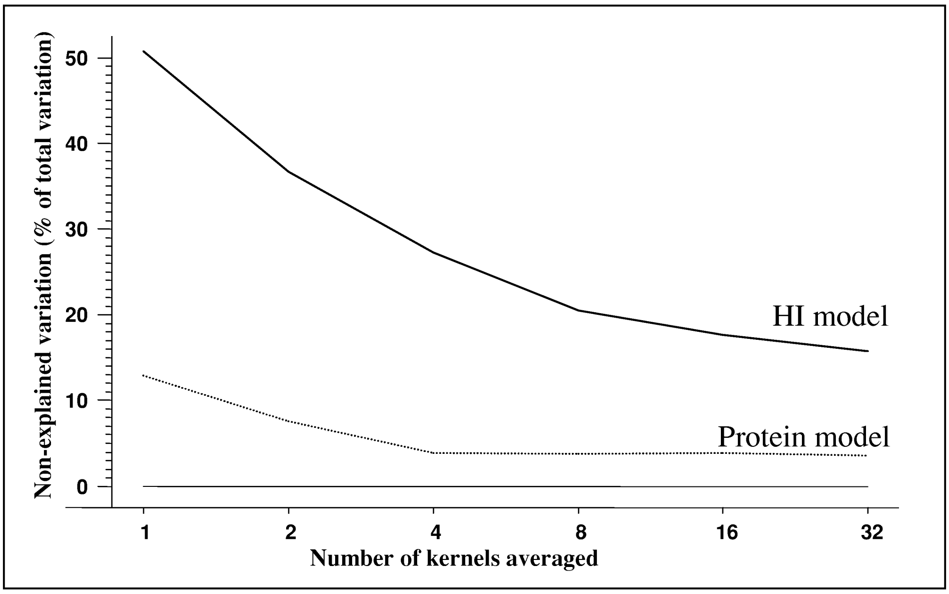


Figure 7. Plot of non-explained variation in percent of total variation versus levels of averaging for the NIT prediction model for hardness (solid line) and protein content (dotted line)

Table III. Correlation coefficients (r) and prediction errors (RMSECV) of the replicate simulation by averaging kernels for the NIT prediction models for Kjeldahl protein content and SKCS hardness

Number of kernels averaged	Protein model		SKCS HI model	
	r	RMSECV	r	RMSECV
1	0.93	0.58	0.70	18.6
2	0.96	0.44	0.80	15.8
4	0.98	0.32	0.85	13.7
8	0.98	0.31	0.90	11.9
16	0.98	0.31	0.91	11.0
32	0.99	0.31	0.93	10.4

Conclusions

By applying a single kernel procedure in which the non-destructive analyses are conducted prior to the destructive ones, several single kernel characteristics can be linked directly to the same functional unit, the single seed, to be used in cereal processing and breeding. In this investigation, the development of non-destructive screening methods for single seed protein content, vitreousness, density and SKCS hardness index for the same set of kernels has been studied by applying this type of procedure.

Table IV. Summary of the non-destructive screening methods on single kernels

Data (**X**)	Parameter (**y**)	r[a]	RMSEP[b]	RE[c]
NIT 850-1050 nm (scatter corrected)	Protein	0.98	0.48	4.7%
NIT 850-1050 nm (raw)	Vitreousness[d]	0.76	4.6	12.6%
NIT 850-1050 nm (raw)	Density	0.63	0.035 [e]	13,4%
GrainCheck data plus kernel weight	Volume	0.99	0.001[e]	2.9%
GrainCheck data plus kernel weight	Density	0.70	0.030 [e]	11.5%
NIT 850-1050 nm (raw)	Hardness	0.59	20.2	15.5%

[a]: r is the correlation coefficient between measured and predicted

[b]: RMSEP is the average prediction error

[c]: Relative error (RE); RMSECV or RMSECV divided by the range (max-min values); reported in percent

[d]: Determined using GrainCheck

[e]: Models are validated using cross-validation and RMSEP should be RMSECV

The results of the non-destructive prediction models for single kernel protein, vitreousness, hardness, volume and density are summarised in Table IV. NIT spectroscopy, in combination with multivariate analysis, shows excellent ability to determine protein content, and only shows some ability for determination of single kernel vitreousness. It is concluded that the non-destructive determination of kernel density, on the other hand, either based on NIT spectroscopy or a combination of kernel weight and image analysis, needs further improvement for practical use.

The use of a true single seed hardness determination, in terms of SKCS HI, as reference values in a NIT prediction model resulted in poor predictability. However, the results shown in Figure 7 and Table III suggest that raw NIT spectra actually contain more information about kernel texture than the poor prediction model in Figure 6 suggests. It seems that a single seed reference method for hardness determination with greater accuracy is needed in order to achieve a good and useful NIT prediction model. If this is possible, there seems to be a potential for the development of a model, which would allow the use of raw NIT spectra for a non-destructive single seed hardness analysis.

For practical use of single seed near infrared spectroscopy as an homogeneity tool, it is important that the measurements are automated, as in the new combined SKCS-NIR instrument (Delwiche and Hruschka, 2000; Dowell et al., 1999). The Infratec 1255 single seed measurements provides excellent single seed protein data that are much easier to obtained than the traditional Kjeldahl method, but the single seed handling is still not automated and the measurements are quite time consuming when analysing high number of kernels. When applied automatically, near infrared spectroscopy on single seeds, alone or in combination with other automated non-destructive techniques, has a great potential as routine homogeneity analysis. This might not only be limited to protein and hardness, but also for other quality parameters in cereals, as the method is used today on bulk samples.

Acknowledgments

The authors are grateful to The Directorate for Food, Fisheries and Agri Business for financial support through the Danish Cereal Network-B4. Karen Thomsen is thanked for development and testing of the single seed Kjeldahl method, and The Danish Institute of Agricultural Sciences for providing the wheat samples. Dr. Bo Löfqvist, BoMill AB, Sweden, is thanked for inspiring discussions regarding the manuscript.

Literature cited

Abe, H., Kusama, T. and Kawano, S. 1995. Non-destructive determination of protein content in a single kernel of wheat and soybean by near infrared spectroscopy. In: A. M. C. Davies and P. Williams (Eds.), Near Infrared Spectroscopy: The Future Waves. NIR Publications, Chichester, UK, pp. 457-461.

Berman, M., Bason, M. L., Ellison, F., Peden, G., and Wrigley, C. W. 1996. Image analysis of whole grains to screen for flour-milling yield in wheat breeding. Cereal Chem. 73(3):323-327.

Chtioui, Y., Bertrand, D., Dattée, Y. and Devaux, M.-F. 1996. Identification of seeds by colour imaging: Comparison of discriminant analysis and artificial neural network. J sci food agr 71:433-441.

de Noord, O. E. 1994. The influence of data preprocessing on the robustness and parismony of multivariate calibration models. Chemom Intell Lab Systems 23:65-70.

Delwiche,S.R., 1993. Measurement of single-kernel wheat hardness using near-infrared transmittance. Trans. ASAE 36(5):1431-1437.

Delwiche, S. R., 1995. Single wheat kernel analysis by near-infrared transmittance: protein content. Cereal Chem. 72(1):11-16.

Delwiche, S. R. 1998. Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. J cereal sci 27: 241-254.

Delwiche, S. R. and Hruschka, W. R. 2000. Protein content of bulk wheat from near-infrared reflectance of individual kernels. Cereal Chem. 77(1):86-89.

Delwiche, S. R. and Massie, D. R. 1996. Classification of wheat by visible and near-infrared reflectace from single kernels. Cereal Chem. 73(3):399-405.

Dowell, F. E. 2000. Differentiating vitreous and nonvitreous durum wheat kernels by using near-infrared spectroscopy. Cereal Chem. 77(2):155-158.

Dowell, F. E., Ram, M. S. and Seitz, L. M. 1999. Predicting of scab, vomitoxin, and ergosterol in single wheat kernels using near-infrared spectroscopy. Cereal Chem. 76(4):573-576.

Geladi, P., MacDougall, D. and Martens, H. 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Appl. Spectroscopy 39(3):491-500.

Martens H. and Næs T. 1993. Multivariate Calibration. Wiley, New York.

Martin, C., Rousser, R. and Brabec, D. 1993. Development of a single-kernel wheat characterization system. Trans. ASAE 36:1399-1404.

Orman, B. A. and Schumann, R. A. 1992. Nondestructive single-kernel oil determination of maize by near-infrared transmission spectroscopy. JAOCS 69(10):1036-1038.

Patrick, B. E. and Jolliff, G. D. 1997. Nondestructive single-seed oil determination of meadowfoam by near-infrared transmission spectroscopy. JAOCS 74(3):273-276.

Pedersen, D. K., Martens, H., Nielsen, J. P. and Engelsen, S. B. 2002. Near infrared absorption and scattering separated by Extended Inverted Signal Correction (EISC). Analysis of NIT spectra of single wheat seeds. Applied Spectroscopy, in press.

Pomeranz, Y. and Williams, P. C. 1990. Wheat hardness: Its genetic, structural and biochemical background, measurement and significance. Advances in Cereal Science and Technology 10:471-544.

Wang, D., Dowell, F. E. and Chung, D. S. 2001. Assessment of heat-damaged wheat kernels using near-infrared spectroscopy. Cereal Chem. 78(5):625-628.

Williams, P. C. 1991. Prediction of wheat texture in whole grains by near-infrared transmittance. Cereal Chem. 68(1): 112-114.

Zayas, I., Lai, F. S. and Pomeranz, Y. 1986. Discrimination between wheat classes and varieties by image analysis. Cereal Chem. 63(1):52-56.

Zayas, I., Martin, C. R., Steele, J. L. and Katsevich, A. 1996. Wheat classification using image analysis and crush-force parameters. Trans. ASAE 39(6):2199-2204.

Paper 7

Study of NIR spectra, particle size distributions and chemical parameters of wheat flours: a multi-way approach

Jesper Pram Nielsen, Dominique Bertrand, Elisabeth Micklander,
Phillip Courcoux and Lars Munck

Study of NIR spectra, particle size distributions and chemical parameters of wheat flours: a multi-way approach

J. Pram Nielsen,[a] D. Bertrand,[b] E. Micklander,[a] P. Courcoux[b] and L. Munck[a]

[a]*Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C., Denmark*

[b]*Unité de Sensométrie et de Chimiométrie, ENITIAA/INRA, rue de la Géraudière, BP 44322 Nantes Cedex 03, France*

Near infrared (NIR) reflectance spectra contain information about both physical and chemical characteristics of flour samples and have great potential for on-line/at-line quality control in a flour mill. The addition of physical characteristics such as particle size distribution data to the NIR spectra and chemical composition data of wheat flour samples was anticipated to provide a better understanding and translation of multivariate measurements into the operational routines and experiences of mill operators. This was studied using a multi-way model called "Analysis of Common Dimensions and Specific Weights" (COMDIM). By this method the underlying dimensions across several data tables with different numbers of variables are defined and the scores and loadings are interpretable in the same way as in a classical Principal Component Analysis. The method was applied on raw NIR spectra as well as after correcting the NIR spectra using the Standard Normal Variate (SNV). The model output in terms of weights, scores and loadings were highly interpretable and in agreement with common characteristics of wheat flour samples. Four underlying dimensions explained 99.4% of the total variation, both when analysing raw and SNV-corrected spectra. A comparison of the two analyses clearly shows that correcting the spectra puts more emphasis on the chemical information in the spectra. However, even corrected NIR spectra contain considerable information about the particle size properties of the flour samples. It is suggested that the COMDIM model can be a useful tool in the process control in a flour mill and it can be used on a wide range of multi-way data problems to assure a high degree of interpretability.

Keywords: wheat flours, NIR spectra, particle size, chemical quality, multi-way analysis

Introduction

Industrial dry-milling of wheat consists of a complex procedure of consecutive steps of grinding and size separation. The aim is to obtain a high yield of endosperm flour without contamination of bran particles. The texture of the grain endosperm, i.e. hard or soft, strongly influences the ease of processing the wheat. However, during the milling, adjustments of the different milling devices can be made by the operator in order to improve the yield and purity of the flour. Since the ash content of the bran is considerably higher than in the endosperm, the ash content of the flour is normally used as index of bran contamination.[1] In addition to the purity, the particle size distribution of the different flour outlets is an important parameter to be controlled during the milling process. Several methods are available for measuring the particle size distribution, of which sieving and laser diffraction are among the most commonly used. The two techniques were compared by Nathier-Dufour *et al.*[2] The advantage of the laser diffraction method is that very fine particles (less than 1 µm) can be measured and additionally the method can be installed on-line.

The wheat flour quality is, however, not only limited to the purity and the particle size distribution, it is also defined by other chemical parameters which characterise, for example, good baking quality, storage ability etc. Accordingly, chemical parameters such as moisture content, protein content, fibre content, fat content, starch content and the amount of damaged starch are important for the end product quality. The miller thus has to adjust the different milling machines and mix flour outlets from different milling machines, in order to fulfil these quality requirements as well.

This process control is a multivariate task, since a minor adjustment on one milling machine will have an impact not only on the chemical composition of the flour produced, but also on its physical characteristics and yield. For example, too fine flour has a poor flowability, i.e. it is difficult to transport and handle.

NIR spectroscopy has become a widely used method in analysis of cereals and cereal based products. As reviewed by Osborne *et al.*,[3] determination of quality of wheat flour using NIR has been extensively investigated. In addition, NIR spectra of flours contain information about the particle size, as investigated by Chapelle *et al.*,[4] and the method is currently used as a reference method for wheat endosperm hardness (AACC Method 39-70A).

Since NIR analyses are rapid and non-destructive and the spectra are fingerprints of the physical and chemical properties of the flour samples, this method has great potential as an at-line/on-line monitoring and process control in the milling industry. In the current work, the relationship between NIR spectra, particle size distributions and chemical properties of the samples was investigated by using a multi-way data analysis. In a classical multivariate data analysis, data can be arranged in a two-way matrix with the samples (objects) as rows and the measured variables as columns. The data structure thus has two modes, a sample mode and a variable mode. A two-way data structure can be extended to a three-way data array, where the same variables are measured on the same samples under different circumstances, thus including a third mode, for example, time or temperature. In this way the data structure is no longer a two-way data matrix, but a three-way data array. For a thorough description of multi-way analysis in the field of chemometrics, see Bro.[5] Multi-way analysis has been used to interpret fluorescence spectra[6] and is used in sensometrics.[7] Only few publications have reported the use of multi-way data analysis involving NIR data. Recently, Allosio *et al.*[8] used the PARAFAC algorithm[9] on NIR spectra collected during the processing of barley into malt.

In most multi-way methods the size of a slice of the data array must be the same, i.e. the number of rows and columns at all the measured points in the third mode must be the same. In the present work, this is not the case. In fact, only the sample mode is the same, since the third direction consists of different types of measurements, in which the first data slice is NIR spectra, the second slice is laser particle distributions and the third slice is chemical data, all with different numbers of variables. The objective of this study is therefore to apply a qualitative multi-way method allowing different number of variables in the third mode on NIR spectra, particle size distribution and chemical data tables of wheat flour samples.

Experimental

Sample collection

During full-scale milling of wheat (variety Ritmo), a representative sample of each of six flour outlets was collected (labelled 1–6) together with one sample containing a proportional mix of the six flours (labelled 7). Each of these seven samples was then separated into six sub-samples according to particle size using laboratory equipment. The coarse fraction labelled “a” was separated on a 150 µm sieve using a JEL Laboratory Sifter (JEL, Ludwighafen, Germany). The throughs from the JEL sifter were sifted at 70 µm on an Alpine Air Jet

Sifter (Model A 200 LS). The fraction labelled "b" with a particle size of 70–150 µm was collected. The throughs were further fractionated on an Alpine Air-Classifier (type 132 MP) at different rate settings, giving the following fractions: "c" was separated at a rate setting of 2.5; "d" was separated at a rate setting of 3.4; "e" was separated at a rate setting of 4.0; "f" was the rest remaining after separation of fraction "e".

This set of samples (42 samples in total) should be seen as a "model flour sample set", mainly constructed to expand the chemical quality and granularity of milling flours, and it should not be seen as a sample set representing all flows of material in a wheat mill.

NIR measurements

The NIR reflectance spectra were recorded using a NIRSystems 6500 (Foss NIRSystems, Silver Spring, MD, USA). The flour samples were loaded in Mini Sample Cup Rings (IH-0307) and placed in a Spinning Sample Module (NR-6506). Spectra were collected in the range of 1100–2500 nm with a resolution of 2 nm, which gives 700 variables for each spectrum. Each spectrum was an average of 16 subscans.

Particle size measurements

The particle size distributions of the flours were analysed using a laser diffraction instrument (Master Sizer IP, Malvern Instruments, Malvern, UK) fitted with a lens of focal length 300 mm. Under these conditions, the range of measurement was from 0.5 to 492.5 µm. The particle size intervals were defined by the manufacturer and respected a logarithmic scale, yielding 32 data points. In order to avoid aggregates of particles, water was used as a medium for immediate dispersion prior to measurements. Measurements were achieved in duplicates, and the averages were assessed for further analysis. The volume of each particle was calculated using the diameter by assuming the particles to be spherical in shape. The particle size distributions were expressed as the volume proportions (in percentage) in each class of particle size. The total volume of each distribution was set equal to one.

Chemical analysis

Due to lack of material, a full chemical analysis could not be performed on three of the 42 samples. The following chemical analyses were therefore only performed on 39 of the 42 flour samples. Ash content was estimated using the ICC-standard no. 104/1 method, dry matter using the AACC Method 44-15A, protein content by the Kjeldahl method, starch content using the method of Åman *et al.*[10] and damaged starch using the AACC Method 76-30A.

Mathematical methods

Often, the same collection of samples is studied using various physical methods, which produce independent measurements. In such situations, the results can be gathered in data matrices, in which the rows represent the samples and the columns the measured variables. Let m be the number of data matrices that have the same number of rows n. The complete collection is then represented by a series of matrices, $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_m$. In the more general case, the number of columns $p_1, p_2, ..., p_m$ of these matrices is not identical. The most well-known method for processing such data is INDSCAL, proposed by Carroll and Chang.[11] Qannari *et al.*[12] have proposed a new version of INDSCAL called "Analysis of Common Dimensions and Specific Weights", which we will refer to as COMDIM.

Throughout this presentation it is assumed that the matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_m$ are centred and normalised in order to have equal sums of squares. The purpose of COMDIM is to summarise the information of the different data matrices by a set of vectors of n elements (scores) $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3 ..., \boldsymbol{q}_r$ representing the common underlying dimensions. The basic idea of COMDIM is to consider each association matrix such as $\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$ rather than the matrix $\boldsymbol{X}_i$ itself. The matrix $\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$ reflects similarities among the rows of matrix $\boldsymbol{X}_i$. In COMDIM, a first modelling of matrix $\boldsymbol{X}_i\boldsymbol{X}_i^{\mathrm{T}}$ is expressed as $\lambda_1^{(i)}\boldsymbol{q}_1\boldsymbol{q}_1^{\mathrm{T}}$. The weight $\lambda_1^{(i)}$ is specific both for the dimension 1 and for table i. Similarly, another association matrix, $\boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}}$, corresponding to matrix $\boldsymbol{X}_j$ will therefore be approximated by $\lambda_1^{(j)}\boldsymbol{q}_1\boldsymbol{q}_1^{\mathrm{T}}$. Figure 1 shows a graphical representation of COMDIM.

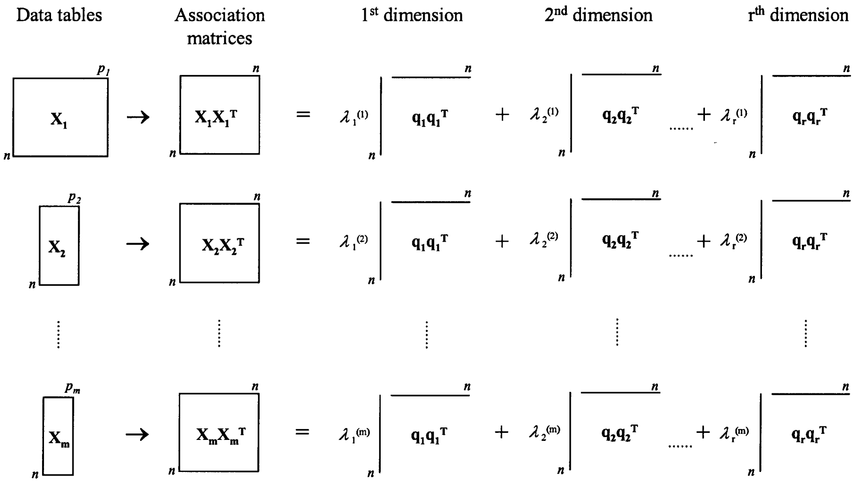


Figure 1. Graphical representation of COMDIM.

In order to estimate the weights $\lambda_1^{(1)}$, $\lambda_1^{(2)}$, ..., $\lambda_1^{(m)}$ and the (single) component q_1, it is necessary to define a criterion of quality, represented by a loss function L_1. It seems logical to minimise the sum of squares of the differences between estimated and observed association matrices. L_1 is therefore expressed as:

$$L_1 = \sum_{i=1}^{m} \left\| X_i X_i^T - \lambda_1^{(i)} q_1 q_1^T \right\|^2$$

Qannari *et al.*[12] have given a simple algorithm (not presented here) based on an alternating least square procedure, making it possible to estimate the weights $\lambda_1^{(i)}$ and the component q_1.

The normalised vector q_1 represents the first common dimension. We can now consider the information in the data set that is not already in the first component q_1. This information is estimated by the residuals from the regression of all variables in X_i. As q_1 is normalised to one, the residual $X_i^{(2)}$ is estimated by $X_i^{(2)} = X_i - q_1 q_1^T X_i$.

The matrices of residuals $X_1^{(2)}$, $X_2^{(2)}$, ..., $X_m^{(2)}$ contain the information which was not taken into account by the first component q_1. A second component q_2 can therefore be found by replacing X_1, X_2, ..., X_m with $X_1^{(2)}$, $X_2^{(2)}$, ..., $X_m^{(2)}$. Once q_2 has been found, the second loss function (to be minimised) is obtained from:

$$L_2 = \sum_{i=1}^{m} \left\| X_i^{(2)} X_i^{(2)T} - \lambda_2^{(i)} q_2 q_2^T \right\|^2$$

This leads to the calculation of q_2 and the weights $\lambda_2^{(i)}$, with $i = 1, 2, ..., m$. The same procedure can be reiterated, in order to estimate other components q_3, q_4, ..., q_r. The components are constructed to have norm equal to one, and are orthogonal to each other. The number of relevant dimensions can be appreciated by considering the residuals associated with each component.

It can be useful to calculate loadings that have the same spectral interpretation as the ones obtained in Principal Component Analysis (PCA) or in Partial Least Squares (PLS) regression. They are calculated as the covariance between each component q_j and the columns of the original matrices:

$$v_j^{(i)} = X_i^T q_j$$

It is easy to see that from the components (scores) q_1, q_2, q_3 ..., q_r and the corresponding loadings $v_1^{(i)}$, $v_2^{(i)}$, $v_3^{(i)}$..., $v_r^{(i)}$ an approximation of X_i is given by:

$$X_i = \sum_{j=1}^{r} q_j v_j^{(i)T} + E$$

For each dimension, there are as many loadings as the number of data matrices. If the original data table represents digitised curves such as spectra, it is common practice to represent the loadings as "spectral profiles", emphasising the continuous nature of the data.

The COMDIM method basically gives three main outputs: components q_j, weights $\lambda_j^{(i)}$ and loadings $v_j^{(i)}$ with $i = 1, ..., m$ (number of data tables) and $j = 1, ..., r$ (number of dimensions in the model). The components and loadings are interpretable in the same way as in PCA: the scatter plot of components q_j and q_k reflects the similarities of the observations according to the corresponding components, taking into account all the data matrices. The loadings emphasise the importance of each variable. Both positive and negative values of the loadings can be interpreted. For each component q_j, the weight $\lambda_j^{(i)}$ gives the importance of the data table i in the construction of the dimension j.

In this study, the method described above is applied to three data matrices, corresponding to NIR spectra, laser particle size and chemical data. These tables have the same number of rows ($n = 39$) and the number of columns are 700, 32 and 5, respectively. Initially, COMDIM was applied directly on these three data matrices. Using the raw NIR spectra gives a large importance to the particle size effect. In order to give a larger role to the chemical information that is present in the NIR spectra, the COMDIM was also applied after correcting the NIR spectra using the Standard Normal Variate (SNV) method, as described by Barnes *et al.*[13]

Matlab version 5.3 (MathWorks, Inc.) was used for the data analysis.

Results and discussion

Data collection

The results of the five chemical analyses are shown in Table 1. The flour samples are seen to have a large variation in the amounts of ash, protein and starch as well as in the amount of damaged starch, which fits the purpose of the design of the trial. A considerable variation is, however, also seen in the dry matter content. This can be caused by uneven distribution of water within the wheat kernels as well as by drying during the milling and fractionation. The samples also show a large variation in particle size distribution. Figure 2 shows the laser diffraction results of the flour samples as the average curves of the six size separations. The particle size ranges from a mode around 222 µm in the most coarse fraction (a) to 17 µm in the finest fraction (f).

Table 1. Chemical composition of the 39 flour samples.

Chemical parameter	Mean	Range
Dry matter (%)	88.8	86.6–91.8
Ash (% of DM)	0.56	0.38–1.13
Protein (% of DM)	7.0	3.6–11.8
Starch (% of DM)	83.2	76.0–89.7
Damaged starch (% of DM)	6.8	1.3–17.5

The NIR spectra of the 39 flour samples are displayed in Figure 3 and are presented as raw spectra (a) and SNV-corrected spectra (b). As expected, the raw spectra in Figure 3(a), contain considerable multiplicative scatter mainly due to differences in the average particle size, but are otherwise very sim-

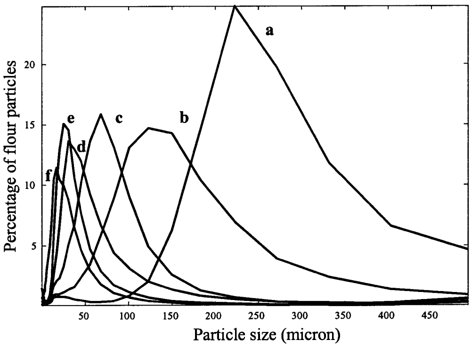


Figure 2. Particle size distributions of the wheat flour samples. "a" to "f" denotes the different particle size fractions with "a" being the most coarse and "f" the finest particles.

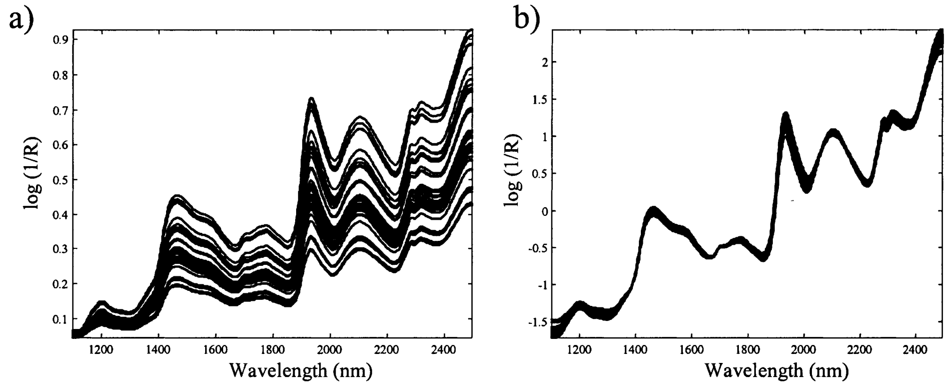


Figure 3. NIR spectra of the flour samples, presented as raw spectra (a) and SNV-corrected spectra (b).

ilar in shape. When corrected for scatter [Figure 3(b)], several well-known absorbance peaks arise corresponding to major chemical constituents of the flour. In general, peaks at around 1450 nm and 1940 nm correspond to the O–H group of for example water.[3] Starch also absorbs at 1450 nm and at 2100 nm and 2460 nm. The absorbance areas of protein are difficult to detect because the protein bands overlap the water and starch.

Multi-way approach

The traditional way of analysing the data set presented here, would be to perform classical multivariate data analysis on the three different data tables separately using methods such as PCA and afterwards try to find the relationships between the three PCA models, or merge the three data tables (with a proper scaling) and perform a PCA on the merged data table. In the multi-way table approach presented here, the three data tables of the flour samples are analysed simultaneously by finding the common dimensions of the three matrices.

Table 2 shows the weights of the analysis of COMDIM using the raw NIR spectra together with the particle size distributions and chemical data. For a given data table, the weights of all possible dimensions sum to one, so a weight for a given dimension within a data table can be considered as percent explained variation. Thus, as seen in Table 2, the first dimension explains 99.44% of the total variation in the NIR spectra, 46.84% of the variation in the particle size data and only 21.22% of the variation in the chemical data. In COMDIM the common components are calculated in order to explain as much vari-

Table 2. Table of the weights of the analysis of common dimensions using raw NIR spectra.

Dimensions →	1	2	3	4
Tables ↓				
NIR (raw) data	0.9944	0.0011	0.0007	0.0016
Particle size data	0.4684	0.0577	0.2860	0.1152
Chemical data	0.2122	0.5860	0.0110	0.0117
Explained variation (%)	80.5	15.4	3.0	0.5

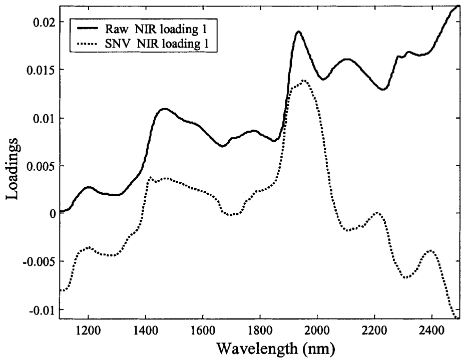


Figure 4. First spectral loading of COMDIM using raw NIR spectra (——) and SNV-corrected NIR spectra (· · ·).

ation as possible of all the three matrices combined; here the first component explains 80.5% of the total variation in the three data tables. The scatter in the NIR spectra mainly causes this first dimension, and this dimension can therefore be considered as an average particle size dimension. The NIR loading (Figure 4, solid line) of the first dimension verifies this nicely, since the loading is almost equal to a mean spectrum of the 39 flour samples.

From Table 2 it is further seen that the second dimension mainly explains the chemical data (58.60%) and only contributes slightly to the description of the NIR and the particle data. The third and fourth dimensions, only explaining 3.5% of the total variation, are almost not influenced by the spectral or chemical data, but describe 28.6% and 11.52% of the particle size variation, respectively. Since it is considered that the first dimension mainly describes the average particle size, these components might describe differences in the particle size distributions.

It is well known that changes in particle size cause a change in the scatter of light. In order to investigate the effect of spectral scatter correction an analogue analysis of COMDIM was performed using the SNV-corrected spectra. Table 3 shows the weights of this analysis. The first dimension now comprises 78.6% of the total variation, which is nearly the same as for raw NIR spectra, but the explained variation within the three data matrices has changed considerably. The explained variation of the NIR spectra has decreased from 99.44% using the raw spectra to 86.94% using the SNV-corrected. The explained variation of the particle size data has also decreased, whereas the explained variation of the chemical data has increased considerably from 21.22% to 34.14%. This shows that removal of the light scatter from the NIR spectra reduces the spectral information of the physical characteristics of the samples and increases the weight of spectral information of the chemical properties of the samples. The NIR loading of the first dimension is shown in Figure 4 (dashed line). This loading, as compared to the first loading using raw NIR spectra (solid line), seems to be a combination of scatter and chemical information. In the area of 1100–1600 nm the two curves seem quite similar, mainly representing pure spectral offset, whereas the loading of the SNV-corrected NIR spectra seems to show a more chemically influenced pattern in the range of 1600–2500 nm. A broad overlapping peak between 1920–2000 nm is seen, which is mainly due to the absorption of water.

Table 3. Table of weights of the analysis of common dimensions using SNV-corrected NIR spectra.

Dimensions →	1	2	3	4
Tables ↓				
NIR (SNV-corrected) data	0.8694	0.0691	0.0139	0.0048
Particle size data	0.3188	0.1615	0.3035	0.1268
Chemical data	0.3414	0.4822	0.0030	0.0251
Explained variation (%)	78.6	16.2	4.0	0.6

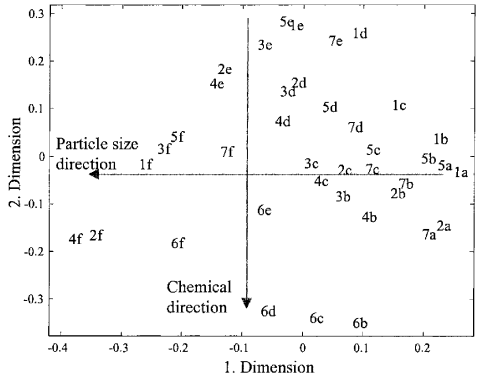


Figure 5. Plot of sample scores of the first and the second dimension using the SNV-corrected NIR spectra. The first digit (1 to 7) refers to the flour outlet and the second digit (a to f) refers to the particle size fraction.

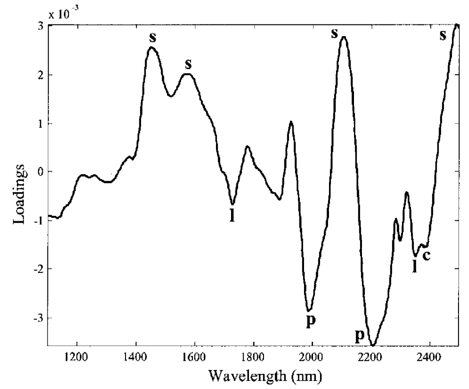


Figure 6. Second spectral loading of COMDIM using SNV-corrected NIR spectra. **s**: absorption of starch; **p**: absorption of protein; **l**: absorption of lipids; **c**: absorption of cellulose.

The weights of the second dimension (Table 3) show that this is mainly a chemical dimension, although there is some contribution from the particle size distribution and a minor contribution from the spectra. As in the analysis based on raw NIR spectra, the third and fourth dimensions describe the particle size distributions.

The weights of the two analyses of COMDIM show that SNV gives patterns that would be expected when comparing the use of raw and SNV-corrected spectra. The COMDIM on the SNV-corrected spectra gives higher weights and more interpretable chemical information and only these results will be discussed in the following.

Figure 5 shows the sample scores of the first and second dimension, which are conceptually analogous to the scores in a PCA sense. The first dimension shows a clear distribution according to particle size fraction going from "a" to "f" when moving from right to left in the plot. By considering the second dimension and consulting the raw chemical data table (not shown), a nice distribution according to chemical properties is seen, when moving from the upper part of the plot with samples having high starch, low protein and low ash content to the lower part of the plot with samples having low starch, high protein and high ash content. This dimension explains 48.22% of the variation in the chemical data, but only 6.91% of the spectral variation. The corresponding NIR loading, shown in Figure 6, can be easily interpreted with regard to chemical composition. Intense positive peaks corresponding to the absorbance of starch are seen at approximately 1450 nm, 1575 nm, 2100 nm and 2490 nm. Negative peaks seen around 1980 nm and 2200 nm are due to protein absorption and some minor negative peaks at 1725 nm and 1880 nm and in the range of 2330–2360 nm are also visible. These peaks are mainly due to absorbance of cellulose and lipids, which are substances at high concentration in the outer part of the endosperm and in the bran of the kernel, and are therefore indicative of bran contamination and high ash content in the flour.

This spectral NIR loading of the second dimension can be further verified by looking at the chemical loading. The chemical loadings for the first and second dimensions are shown in Figure 7. These are presented as bars. The second loading (shown in grey) has a high positive value for starch content and negative values for both ash and protein content which fully agrees with the spectral loading of the second dimension (Figure 6) and complies with common knowledge regarding milling of wheat. The dry matter content and the amount of damaged

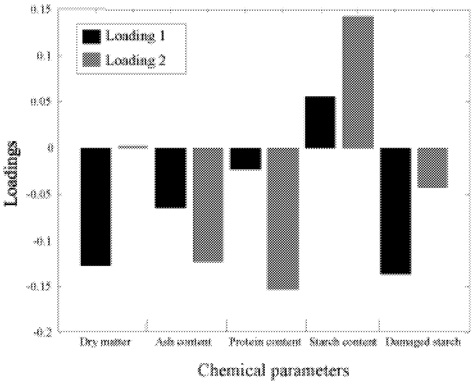


Figure 7. First and second chemical loadings of COMDIM using SNV-corrected NIR spectra.

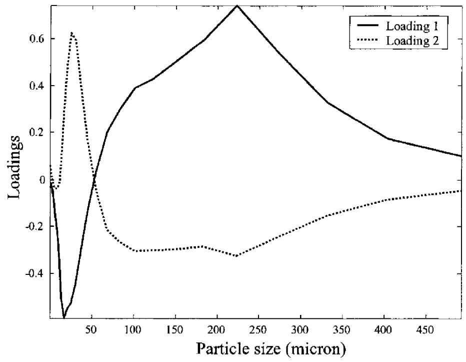


Figure 8. First (—) and second (· · ·) particle size loadings of COMDIM using SNV-corrected NIR spectra.

starch only seem to have minor influence on the second dimension; however, these two parameters seem to have great influence on the first dimension (loading shown in black). In addition, these two variables seem to be correlated. The first dimension is mainly thought to describe the average particle size, even though the NIR spectra have been SNV-corrected, as shown in Figure 5. This dimension, however, also explains 34.14% of the chemical variation, which seems mainly to be founded in dry matter and damaged starch. This might be due to the fact that the average particle size, the dry matter content and the content of damaged starch are correlated: the smaller the particles, the more damaged starch and the higher the content of dry matter. The NIR loading of the first dimension (Figure 4, dashed line) might then be considered as the loading for these two chemical parameters; however, it could also be based on internal correlations.

Thus so far, the interpretation has been focused on the spectra and the chemical loadings, but when using this multi-table approach, it is possible to consider the corresponding particle size dimensions. These two loadings are shown in Figure 8. The first loading (solid line) nicely reflects the average particle size curve (using centred data) having a negative peak at 17 µm corresponding to the mode of the fine particles and a positive peak around 222 µm representing the mode of the coarse particles. In the score plot (Figure 5), the first dimension is expanded by the samples "f" in the left part of the plot having fine particles, and the samples "a" in the right part of the plot having coarse particles. The second particle size loading, explaining only 16.15% of the particle size variation shows a clear positive peak at approximately 27 µm. This means that particles around this size have a high amount of the chemical profile shown in the second chemical loading in Figure 7 (grey), thus having a high starch content and a low protein and ash content. These relatively fine, high-starch particles are mainly the samples "d" and "e" in the upper part of the score plot in Figure 5, where the samples labelled "6" in the lower part of the plot are the most impure (highest in ash content).

When considering dry milling from a practical point of view, the aim is to produce a pure high-yielding flour with a given quality and still keep the particle size as large as possible, in order to secure a high flowability in the transportation and sieving devices. The miller must therefore find a compromise between the chemical properties of the flour and the flour particle size when operating and optimising the mill. The location of a given sample in the score plot in Figure 5 indicates whether this sample has the right quality and particle size properties. When optimising the mill, newly acquired NIR spectra of the flour outlets could be projected onto the COMDIM NIR loadings and the particle size and chemical properties of the flour sample could be evaluated only on the basis of the location (or movements) in the score plot.

Conclusion

The COMDIM was applied on three data tables representing NIR spectra, particle size data and chemical data of wheat flour samples. The output from this analysis that conceptually can be seen as a PCA across several data tables has been useful in interpreting the relationship between the three data tables. In this way, the importance of the information from the different flour measurements has been assessed and interpreted in a more straightforward manner than doing PCA on three data tables separately. The analyses showed that both when using raw and SNV-corrected NIR spectra 99.4% of the total variation was described in only four underlying dimensions. A comparison of the two analyses clearly shows that SNV-correcting the spectra puts more emphasis on the chemical information in the spectra. However, even corrected spectra contain considerable information on the particle size properties of the samples. The spectral, particle size and chemical loadings of the underlying dimensions were interpretable and showed patterns in agreement with prior knowledge regarding characteristics of wheat flour, and it is therefore concluded that the current multi-way method is applicable for investigating of different types of measurements on the same sample set.

In the present context, this multi-table approach can be used to obtain a better understanding of the relationship between physical and chemical properties of wheat flours by getting a direct link, in terms of loadings, between the two types of data. By comparing the results of the analyses of COMDIM performed on different pre-treated spectra, the effect of spectral pre-treatments can be analysed. The current data analysis can, of course, be applied on more or fewer data tables, for example the NIR and particle size data, or the particle size and chemical data in order to emphasise on more specific relationships.

Data analysis as presented here could be a useful tool when NIR spectroscopy is applied as a quality control in a flour mill. Normally, acquired NIR spectra are used in a quantitative way as in predicting the chemical constituents using PLS regression. Instead of doing so, it could be possible to perform a "qualitative calibration" where the flour samples are ranked according to their NIR score values, or a combination of several scores, for example if the miller prefers samples having a high value of the first score and a low value of the second score. Such a qualitative calibration can, of course, also be done by a classical PCA on the NIR spectra, but the advantage of this multi-way approach is that the variation in both the particle size and chemical data is used simultaneously to guide the decomposition of the spectral data. The miller can use this to optimise for different quality parameters.

Acknowledgements

The authors wish to thank the Cascade EU-FAIR 96-1105 project for financial support of this co-operative work between Food Technology, The Royal Veterinary and Agricultural University, Copenhagen and INRA, Phytec, Nantes. We are very grateful to Professor E.M. Qannari (Enitiaa/INRA-Nantes, France) for hosting J. Pram Nielsen in Nantes during finalisation of the manuscript and to Dr Bo Löfqvist, BoMill AB, Sweden for constructive discussions concerning the manuscript.

References

1. J. Abecassis, in *Primary Cereal Processing*, Ed by B. Godon and C. Willm. VCH Publishers, New York, p. 291 (1994).
2. N. Nathier-Dufour, L. Bougeard, M.F. Devaux, D. Bertrand and F. Le Deschault de Monredon, *Powder Technology* **76**, 191 (1993).
3. B. Osborne, T. Fearn and P.H. Hindle, *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. Longman Scientific and Technical, Harlow, UK (1993).
4. V. Chapelle, J.P. Melcion, P. Robert, D. Bertrand, *Sciences des Aliments* **9**, 387 (1989).

5. R. Bro, *Multi-way analysis in the food industry. Models, Algorithms and Applications*. Doctoral thesis, University of Amsterdam, The Netherlands (1998).
6. C.A. Andersson, *Exploratory Multivariate Data Analysis with Applications in Food Technology.* PhD Dissertation, The Royal Veterinary and Agricultural University, Frederiksberg, Copenhagen, Denmark (2000).
7. P.M. Brockhoff, D. Hirst and T. Næs, in *Multivariate Analysis of Data in Sensory science*, Ed by T. Næs and E. Risvik. Elsevier Publishers, pp. 307-342 (1996).
8. N. Allosio, P. Boivin, D. Bertrand and P. Courcoux, *J. Near Infrared Spectrosc.* **5**, 157 (1997).
9. R. Bro, *Chemom. Intell. Lab. Systems* **38**, 149 (1997).
10. P. Åman, E. Westerlund and O. Theander, "Determination of starch using a thermostable a-amylase", in *Methods in Carbohydrate Chemistry: Enzymic Methods*, Ed by J.N. Miller, D.J. Manners and R.J. Sturgeon. John Wiley & Sons, vol X, pp. 111–115 (1994).
11. J.D. Caroll and J.J. Chang, *Psychometrika* **35**, 283 (1970).
12. E.M. Qannari, I. Wakeling, P. Courcoux and H.J.H. MacFie, *Food Quality and Preference* **11**, 151 (2000).
13. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc.* **43**, 772 (1989).

Received: 9 April 2001
Revised: 12 September 2001
Accepted: 13 September 2001
Web Publication: 28 November 2001

Appendix 1

Interval partial least squares regression (iPLS): A comparative chemometric study with an example from near infrared spectroscopy

Lars Nørgaard, Arild Saudland, Jesper Wagner, Jesper Pram Nielsen,
Lars Munck and Søren Balling Engelsen

Interval Partial Least-Squares Regression (*i*PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy

L. NØRGAARD,* A. SAUDLAND, J. WAGNER, J. P. NIELSEN, L. MUNCK, and S. B. ENGELSEN

The Royal Veterinary and Agricultural University, Food Technology, Chemometrics Group, Department of Dairy and Food Science, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

A new graphically oriented local modeling procedure called interval partial least-squares (*i*PLS) is presented for use on spectral data. The *i*PLS method is compared to full-spectrum partial least-squares and the variable selection methods principal variables (PV), forward stepwise selection (FSS), and recursively weighted regression (RWR). The methods are tested on a near-infrared (NIR) spectral data set recorded on 60 beer samples correlated to original extract concentration. The error of the full-spectrum correlation model between NIR and original extract concentration was reduced by a factor of 4 with the use of *i*PLS ($r = 0.998$, and root mean square error of prediction equal to 0.17% plato), and the graphic output contributed to the interpretation of the chemical system under observation. The other methods tested gave a comparable reduction in the prediction error but suffered from the interpretation advantage of the graphic interface. The intervals chosen by *i*PLS cover both the variables found by FSS and all possible combinations as well as the variables found by PV and RWR, and *i*PLS is still able to utilize the first-order advantage.

Index Headings: **Interval PLS; Variable selection; NIR, Principal variables; Forward stepwise selection; Recursively weighted regression; Beer; Extract.**

INTRODUCTION

Full-spectrum regression methods such as partial least-squares regression (PLS) and principal component regression (PCR) have abundantly documented their efficiency within the development of rapid spectral analytical screening methods.[1,2] We have previously applied this approach in exploratory spectral investigations of sugar,[3] pectins,[4] and frying oils[5] employing fluorescence and near-infrared (NIR) as well as Fourier transform infrared and Fourier transform Raman spectroscopy. Chemometricians and data analysts are familiar with the concepts and often favor the use of principal components or latent variables as these aim to represent global orthogonal noncorrelated data structures deduced from the highly intercorrelated spectral ensembles. Spectroscopists, on the other hand, usually have a preference for variables or intervals of variables in the original variable space because these represent interpretable chromophores, fluorophores, or vibraphores and because a strict orthogonal decomposition is not realistic. Other important reasons for the development of methods for spectral variable or interval selection are the improvement of models with respect to predictive ability and the possibility for development of very fast instruments including reduction of the production costs for such instruments by employing a few critical regions; e.g., in a filter instrument. A short time of analysis makes the instruments suitable for rapid on-line measurements; e.g., within the area of process monitoring and control. With respect to data reduction, variable selection may also be a realistic method since spectral data contain a high degree of covariance and, as such, large amounts of redundant information. The need for chemometric methods for variable or interval selection where information is optimally preserved is therefore very large.

One of the main advantages in multivariate data analysis and latent variable methods is the possibility of projecting multivariate data into few dimensions in a graphical interface. We propose a new type of graphical output which will enhance the information content for standard multivariate regression methods such as PCR or PLS. The method that we will focus on is a new graphically oriented approach for local regression modeling of spectral data called interval partial least-squares regression (*i*PLS). An NIR spectral data set is investigated, which has proven to give suboptimal solutions in standard full-spectrum PLS applications. The purpose of the interval and variable selection is to optimize the predictive power of PLS regression models and to aid in interpretation. The investigation has the aim of making a comparative study of the prediction performance of selected different methods for selection of manifest variables or intervals of manifest variables compared to the results based on full-spectrum models. In addition to *i*PLS, the principal variables (PV) method as developed by Höskuldsson,[6] forward stepwise selection (FSS) of variables, and a newly presented method called recursively weighted regression (RWR) will be investigated.[7] The results from using these methods will be compared to results from using full-spectrum PLS results. Common to the methods investigated is that they are based on no, or simple, search strategies. Methods based on intensive heuristic search strategies such as genetic algorithms will not be investigated in this paper.[8] An important contribution to the discussion of spectral variable selection was recently given by Spiegelman et al. in their paper on a theoretical justification of wavelength selection in partial least-squares regression.[9] In the literature, the use of principal variables as an alternative to principal components for a single matrix was presented by McCabe,[10] and this topic was also treated by Krzanowski in a study on how to preserve multivariate data structure using principal components analysis

Received 7 June 1999; accepted 25 September 1999.

* Author to whom correspondence should be sent.

0003-7028/00/5403-0413$2.00/0

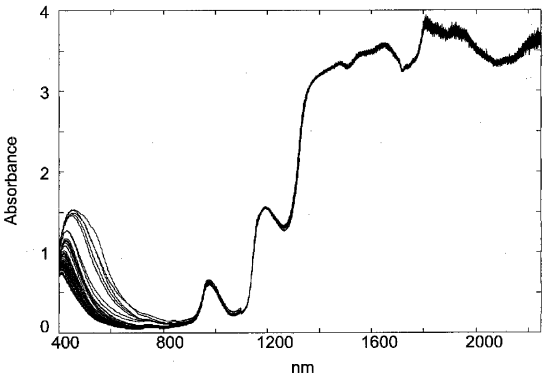


FIG. 1. NIR/visible spectra recorded on 60 beer samples in the wavelength range 400 to 2250 nm; in total 926 variables per sample.

(PCA).[11] Subsequently, Höskuldsson presented a general method capable of dealing with variable selection either in the PCA situation or in the regression situation.[6]

EXPERIMENTAL

Programs. Calculations were performed with Matlab Version 5.2.0 (MathWorks, Inc., Natick, MA) installed with the PLS_Toolbox Ver. 2.0.0b (Wise & Gallagher; Eigenvector Research, Manson, WA) and Unscrambler Version 7.01 (CAMO A/S, Norway). A modified version of a Matlab program made by Agnar Höskuldsson was used for the principal variable selection. The algorithms for *i*PLS, forward stepwise selection, and recursively weighted regression were programmed in the Matlab language by the authors.†

Data Set and Measurement Conditions. We will demonstrate the use and performance of the different variable selection methods by a comparative application to a spectroscopic data set dealing with the determination of the amount of extract from NIR spectra of beers. This data set is an interesting NIR spectral ensemble of 60 beer samples containing a rather large noisy part due to an absorbancy that is too strong, in a region (Fig. 1) dominated by the water component.

Dispersive near-infrared data (including the visual region) at 25 °C were collected with the use of a NIRSystems Inc. (Model 6500) spectrophotometer. The spectrophotometer uses a split detector system with a silicon (Si) detector between 400 and 1100 nm and a lead sulfide (PbS) detector from 1100 to 2500 nm. The NIR/visible transmission spectra were recorded with a 30 mm quartz cell directly on the undiluted degassed beer, and spectral data collected at 2 nm intervals in the range from 400 to 2250 nm were converted to absorbance units.

Original extract concentration is an important quality parameter in the brewing industry, indicating the substrate potential for the yeast to ferment alcohol and serving as a taxation parameter. Original extract concentration was determined by Carlsberg A/S in the range of 4.23–18.76% plato. The data were sorted by extract value, and an independent test set was constructed by selecting every third sample of this full set. There are thus two data sets: one for calibration (40 samples) and one for estimation of prediction error (20 samples). It is assumed that overfitting will be revealed by the independent test set.

CHEMOMETRIC THEORY

Partial Least-Squares Regression. Partial least-squares regression is a predictive two-block regression method based on estimated latent variables and is applied to the simultaneous analysis of two data sets (e.g., spectra and physical/chemical tests) on the same objects[12] (e.g., beer samples). The purpose of the PLS regression is to build a linear model enabling prediction of a desired characteristic ($\mathbf{y}$) from a measured spectrum ($\mathbf{x}$). In matrix notation we have the linear model $\mathbf{y} = \mathbf{Xb}$ where $\mathbf{b}$ contains the regression coefficients that are determined during the calibration step, and $\mathbf{X}$ is the matrix of collected spectra. PLS was first applied to evaluate NIR spectra by Martens and Jensen in 1983,[1] and is now used routinely in academia and industry to correlate (rapid) spectroscopic measurements with related chemical/physical data.

Interval PLS. Interval PLS is an in-house developed interactive extension to PLS, which develops local PLS models on equidistant subintervals of the full-spectrum region. Its main force is to provide an overall picture of the relevant information in different spectral subdivisions, thereby focusing on important spectral regions and removing interferences from other regions. The sensitivity of the PLS algorithm to noisy variables is highlighted by the informative *i*PLS plots.

Interval PLS models are developed on spectral subintervals of equal width, and the prediction performance of these local models and the global (full-spectrum) model is compared. The comparison is mainly based on the validation parameter RMSECV (root mean squared error of cross-validation), but other parameters such as r^2 (squared correlation coefficient), slope, and offset are also evaluated to ensure a comprehensive model overview. Sample and/or measurement abnormalities (outliers) as detected by PLS inner relation plots should generally be removed prior to the application of *i*PLS.

Models based upon the various intervals ($\mathbf{X}_{\text{interval}}$) usually need a different number of PLS components than do full-spectrum models to catch the relevant variation in $\mathbf{y}$. This condition is caused by the varying amount of y-correlated information carried by the interval variables (the larger the spectral interval, the greater the number of substances that are likely to absorb/interfere) and is also related to the noise/interference carried by the variables. However, the selected model dimension has to be common to all the local models in order to make a comparison possible. In order to favor the "best" spectral region, it is natural to let the simplest interval model (i.e., the one with the smallest number of PLS components) guide the selection of the model dimension. A fair comparison of the global and local models requires that the global and local model dimensions be selected separately.

Simple Optimization of the Best Interval from Equidistant iPLS. There is a minimal probability for hitting the

† The *i*PLS algorithm including the optimization module and the NIR data set studied in this work is available from our Web site: http://www.mli.kvl.dk/foodtech/special/specials.htm.

optimal interval with the equidistant subdivisions. A more optimal interval might be found by carrying out small adjustments in the interval limits. The optimization performed consists of the following steps: (1) interval shift; (2) changes in interval width: two-sided (symmetrical), one-sided (asymmetrical, left), or one-sided (asymmetrical, right). Each step is initiated with the optimal interval limits from the previous step. The interval limits are changed one variable at a time and evaluated by the RMSECV provided by application of PLS regression to the interval; this approach works in practice but could be done more elegantly.

Principal Variables. Principal variables is a method for selection of a limited number of original variables (e.g., wavelengths) that describe, as much as possible, the variance in the data matrix (spectra) or, alternatively, covariance in the matrix with a vector with a desired characteristic (chemical/physical measurement).[6] The PV method is initiated by finding the variable (wavelength) that co-varies most with the **y** vector (physical/chemical measurement). This variable is the first principal variable. The original spectral data matrix is then reduced (orthogonalized) with respect to the first principal variable. Then the next covariant variable in the reduced data matrix is selected, and this procedure is followed until the wanted number of principal variables has been calculated. The result of the PV selection is a limited number of the original variables (e.g., wavelengths), while PLS selects latent factors based on information from all original variables. The PV method also works on a single data matrix, in which case the method will search for columns that describe the largest variation; i.e., the method is a general tool for variable selection.

Forward Stepwise Selection of Variables. Forward stepwise selection is a most simple and pragmatic search method in which subsequent variables are selected stepwise by their capability to improve a multiple linear regression (MLR) model. First, all spectral variables are tested individually in univariate linear regression models with extract concentration as the dependent variable. All these models are test set validated, and the variable with the lowest RMSEP (dependent test set) is chosen. Next, all two-variable MLR models are investigated on the basis of the chosen variable in combination with all the remaining variables (one by one). All these models are also test set validated, and the variable that (in combination with the first chosen variable) gives the lowest RMSEP (dependent test set) is chosen. This procedure is continued until the RMSEP (dependent test set) increases by the introduction of a new variable. In the FSS case, a dependent test set is chosen to evaluate the selection of new variables, since an evaluation procedure based on cross-validation leads to severe overfitting.[12,13]

Recursively Weighted Regression. This method is based on an recursive re-weighting of the independent variable block (**X**) by the regression vector **b** calculated from a PLS regression model between **X** and **y**: $\mathbf{x}^i_{n+1} = \mathbf{x}^i_n * b^i_n$, $i = 1$ to number of variables, where b^i_n is the ith element in the PLS regression coefficient vector ($\mathbf{b}_n$) of step number n, and $\mathbf{x}^i_n$ is the ith column of $\mathbf{X}_n$.[7] The algorithm is started with a standard PLS model between $\mathbf{X}_1$ (equal to **X**) and **y**, giving $\mathbf{b}_1$. The re-weighting is repeated 50 times ($n = 1$ to 50) in the calculations presented in this paper in order to ensure that a final solution has been reached. The result is a regression vector $\mathbf{b}_{50}$ that contains only ones and zeros (this binary result is a direct output from the RWR algorithm; i.e., no rescaling of the final regression vector is performed). In the simple case, the number of variables selected (i.e., variables with a corresponding regression coefficient of one) corresponds to the number of latent factors chosen in the original PLS model. This is not the case in more complicated situations. This simple method, which combines multivariate regression and variable selection, has not yet been thoroughly investigated but certainly deserves more attention.

Error Measures. The root mean square error in combination with the correlation coefficient (r) is used as a measure of how a given model performs. RMSE is defined as follows:

$$\text{RMSE} = \sqrt{\frac{\sum (y_{\text{pred}} - y_{\text{ref}})^2}{N}}$$

where y_{pred} is the predicted value, y_{ref} is the laboratory-measured value, and N is the number of samples.

RMSEC is RMSE calculated from the calibration samples, i.e., a measure of fit. RMSECV is calculated from the cross-validated samples, and RMSEP is calculated from the independent test (or prediction) set.[12,13] Correspondingly, r_{cal}, r_{cv}, and r_{pred} are the correlation coefficients for these three situations.

RESULTS AND DISCUSSION

All models are developed on the basis of NIR/visible spectra (**X**) and the response variable extract concentration (**y**). The spectra are shown in Fig. 1. Both mean-centered and autoscaled X data[12] are tested, and all the models developed are validated by segmented cross-validation.[12,13] Five segments are used and they are selected systematically among the 40 calibration samples; i.e., in CV segment number one, the samples 1, 6, 11, 16, 21, 26, 31, and 36 are represented. RMSECV is the parameter governing the variable selection for all tested methods; i.e., the set of variables chosen for a given method is the set that gives the lowest RMSECV among the combinations tested with that method. RMSEP is an estimate of the prediction error based on 20 samples, and its value also reveals whether there are problems with overfitting for some of the methods. In Table I all results on NIR spectral ensemble are compiled.

PLS Full-Spectrum Results. Mean-centered and autoscaled full-spectrum PLS results indicate unstable models when the three RMSE values in Table 1 are compared. Both models are suboptimal, and for the mean-centered model two local minima are seen before the global one at nine PLS components. The improvement in RMSECV when going from five PLS components to nine PLS components is negligible for the model based on autoscaled data (not shown).

***i*PLS Results.** In this section focus is on the situation where the data are autoscaled to provide uniform variance over the entire spectral range (according to the comments made above) and divided into 20 subintervals to show how *i*PLS works. Figure 2 shows y-residual variance characteristics: one for the full-spectrum model and one

TABLE I. Results for NIR on beer samples when using different chemometric methods for variable selection.

Method	Preprocessing	# PCs	# Variables	Interval (nm)	RMSEC	r_{cal}	RMSECV	r_{cv}	RMSEP	r_{pred}
PLS	Auto	9	926	400–2250	0.001	1.000	0.80	0.948	0.40	0.993
	Mean	9	926	400–2250	0.005	1.000	1.31	0.849	0.73	0.961
*i*PLS	Auto	4	46	1228–1318	0.10	0.999	0.15	0.998	0.20	0.997
20 intervals	Mean	4	46	1228–1318	0.12	0.999	0.16	0.998	0.21	0.997
*i*PLS	Auto	3	15	1270–1298	0.24	0.995	0.30	0.992	0.21	0.997
60 intervals	Mean	3	15	1270–1298	0.24	0.995	0.31	0.992	0.22	0.997
*i*PLS	Auto	4	48	1228–1322	0.10	0.999	0.13	0.999	0.18	0.998
optimized[a]	Auto	2	49	1202–1298	0.11	0.999	0.15	0.998	0.17	0.998
PV	Auto	MLR	2	1326, 1184[b]	0.21	0.996	0.24	0.995	0.14	0.999
	Mean	MLR	3	440, 536, 1322[b]	0.41	0.986	0.52	0.977	0.34	0.991
RWR	Auto	2[c]/MLR	2	1184, 1326	0.21	0.996	0.24	0.995	0.14	0.999
	Auto	3[c]/MLR	3	1184, 1320, 1950	0.18	0.997	0.21	0.996	0.15	0.999
	Mean	3[c]/MLR	3	1326, 2234, 2246	0.38	0.988	0.44	0.983	0.37	0.991
FSS	None	MLR	2	1326, 1134[b]	0.16[d]	0.998[d]	0.18[d]	0.997[d]	0.17	0.998
All comb.[e]	None	MLR	2	1128, 1314	0.15	0.998	0.16	0.998	0.20	0.997

[a] Results after optimization based on a 20 interval subdivision.

[b] The variables are found in the written order.

[c] Number of PLS components used in the recursive latent models.

[d] Test set validated with a calibration set of 20 samples and a dependent test set of 20 samples. The RMSECV error corresponds to a dependent test set error, while the RMSEP error is the independent error.

[e] All possible two-variable combinations are tested; in total 428 275 models.

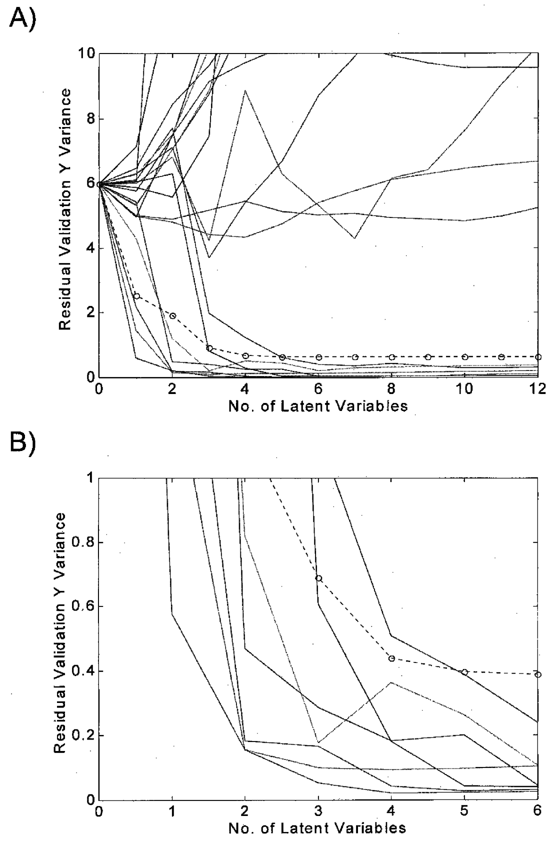


FIG. 2. (**A**) Cross-validated residual y-variance for the full-spectrum model (-o–o) and 20 local models as a function of number of PLS components. (**B**) Enlargement of **A** to show the first minimum at four PLS components for interval 10.

for each of the 20 spectral subdivisions. From Fig. 2 both the local and the global model dimensions are selected. Variance characteristics approaching the abscissa represent promising local models describing most of the systematic variance in the spectral data. In this case there are seven such models, and three of these differ from the rest in having a significant y-residual variance reduction for the very first latent variable. Four PLS components are appropriate for the local models, and in contrast nine PLS components are optimal for the full-spectrum PLS model based on autoscaled data. Figure 3, which demonstrates the central *i*PLS plots, shows expected prediction error (RMSECV) for 20 interval models (bars) and for the full-spectrum model (line) plotted together with a normalized mean spectrum. In Fig. 3A, one PLS component is used in the interval models and for Figs. 3B, 3C, and 3D the number is two, three, and four, respectively. The full-spectrum model (line) is based on nine PLS components in all four plots. It appears from Fig. 3A that only one interval model (number 10) with one latent variable can compete with the full-spectrum model using nine PLS components. However, when two to four (optimal) PLS components are used, several interval models surpass the full-spectrum model.

Interval number 10 (46 variables) was chosen for further optimization: (1) an interval shift of 30 variables to each side was performed, followed by (2) changes in interval width from a chosen minimum of 30 to a chosen maximum of 110 variables [first two-sided symmetrical optimization, then one-sided asymmetrical (left) and one-sided asymmetrical (right) optimization]. The optimization results in an interval in the range 1228–1322 nm (see Table I) with the use of four PLS components. A thorough optimization procedure might include different numbers of PLS components, since a smaller interval might be modeled by a lower number of PLS components. This approach is illustrated by the results from a two-PLS-components solution given in Table I.

Furthermore, the effect of the number of start intervals can be optimized to see how this number influences the

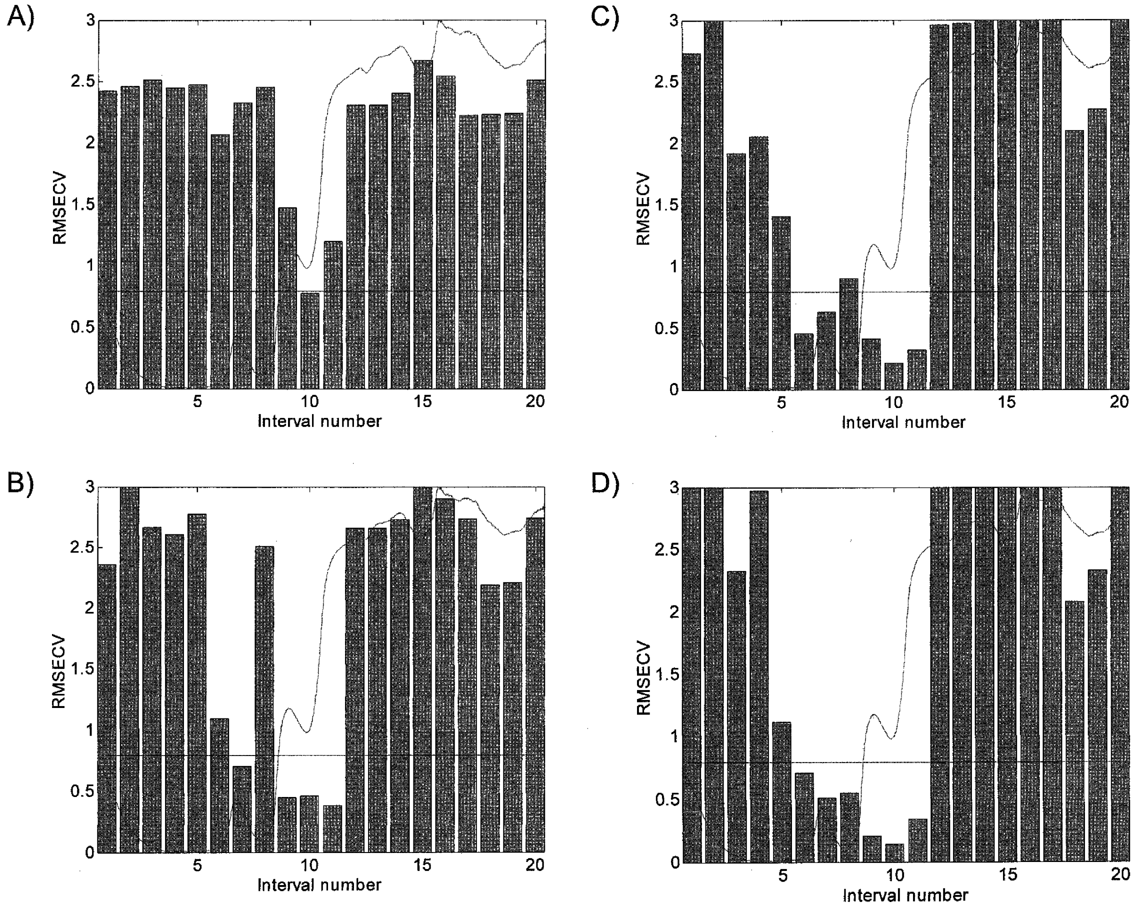


FIG. 3. Cross-validated prediction performance (RMSECV) for 20 interval models (bars) and for full-spectrum model (line) plotted together with the normalized mean spectrum. The interval models used one (**A**), two (**B**), three (**C**), and four (**D**) PLS components, respectively, in the four plots, while the full-spectrum model was kept at nine PLS components (autoscaled data).

results. In Table I results from using both 20 and 60 intervals are shown. The results from using 20 intervals (46 variables in each model) seem more robust than those from the use of 60 intervals (15 variables in each model). Different numbers of intervals can be tested in new applications to see how the information changes with respect to the variables included in the modeling.

Results from Other Methods (PV, FSS, RWR, and All Possible Combinations). Results from principal variables, forward stepwise selection, and recursively weighted regression selection of variables are shown in Table I. Also the optimal result from all possible two-variable combinations (=428 275) are shown. All selected variables are based on the lowest value of RMSECV.

Discussion. Comparing Figs. 1 and 3D, we see that both the noisy region from 1400 to 2250 nm and the systematic visual region from 400 to 800 nm are found to be of no relevance when building correlation models to the original extract. In this way *i*PLS gives an overview of the spectral data and reveals the interesting parts of the spectrum, helping in chemical interpretation. In this case the transparent spectral NIR region between the visual region (400–800 nm) and the NIR region where the strongly absorbing O–H vibrations of the water begin to appear (from 1400 nm and up) holds the predictive performance with respect to the original extract measurement. Except for the second overtone of the O–H stretch at ~ 970 nm, this NIR region is dominated by C–H and N–H stretching overtones. It is seen that this data set without *a priori* knowledge may cause severe troubles for the PLS algorithm. The experienced spectroscopist would remove the noisy spectral region prior to PLS calibration, but this investigation aims at illustrating the usefulness of variable or interval selection when (PLS) calibration is performed on new data to which no prior knowledge is available, or when PLS is applied to data sets which are too large and/or inhomogeneous for standard exploratory PLS investigation.

From Table I we see that none of the full-spectrum models perform well compared to the selection methods.

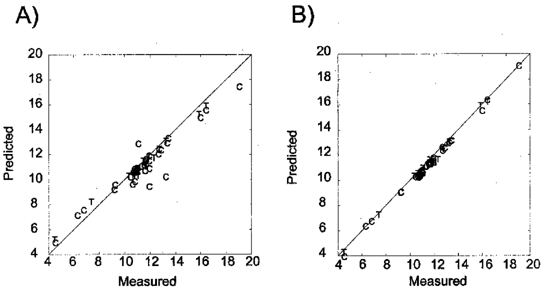


FIG. 4. (**A**) Predicted vs. measured plot for the full-spectrum PLS model with nine PLS components on autoscaled data. (**B**) Corresponding plot for the best interval (1228–1318 nm) from an *i*PLS model with 20 subdivisions without further optimization. In both plots, C denotes cross-validated predictions and T denotes the independent test set predictions.

There is no systematic trend between the RMSEC, RMSECV, and RMSEP values, which is to be expected for a robust model. This observation is due to the rather noisy region in the range 1400–2250 nm, which spoils the full-spectrum PLS model. Also mean-centered PV and RWR fail to find a suitable combination of variables. Compared with all possible two-variable combinations, the methods of *i*PLS, PV, and RWR work well on autoscaled data; *i*PLS also works well on mean-centered data. FSS is scaling independent, and it is the only variable selection method that finds a pair of variables that compares very well to the result found by all possible combinations of two variables. Both the RWR and PV methods select a different combination of variables that is almost alike for the two methods. The intervals chosen by *i*PLS cover both the variables found by FSS and all possible combinations as well as the variables found by PV and RWR. The RWR method is robust with respect to the number of PLS components chosen as a starting point (two or three) for the PLS models when using autoscaling. This observation is reflected in the regression coefficients of the RWR model based on three PLS components resulting in a three-variable multiple linear regression model with the following regression coefficients: $(b_0, b_1, b_2, b_3) = (83.94, 84.06, -88.94, -2.87)$. The regression coefficient b_3 of the noisy variable at 1950 nm has a low numerical value compared to b_0, b_1, and b_2 when considering the absorbance values at the three variables, indicating that the first two variables (1184 and 1320 nm) are sufficient for building a regression model; b_0 is the offset in the regression model.

The optimized PLS model does not give a decrease in RMSECV, reflecting that it is difficult to optimize the actual model further. The decrease from 0.15 to 0.13 is not significant when we take into account the uncertainty of the original extract measurement (estimated to be ~ 0.02–0.04% plato). This observation is supported by the fact that the optimized interval (1228–1322 nm) is almost exactly the same as the first chosen interval (1228–1318 nm), reflecting a chance improvement. Furthermore, by using the interval 1202–1298 it is possible to obtain comparable RMSECV results with only two PLS components.

In Fig. 4 the predicted vs. measured plots for a full-spectrum model and the best interval model (without optimization) are given to show the significant decrease in RMSECV and RMSEP. In this case the interval model is superior to the full-spectrum model. It should be stressed that knowledge of the reproducibility of the spectral measurements (not available here) might be used in preprocessing of the spectral data so that the noisy part of the spectrum is down-weighted in a PLS analysis.

Furthermore, it should be emphasized that especially minimalistic variable selection reduces the power of multivariate outlier control and increases the influence of spectral noise. The trade-off between the measurement of few variables and a reduced quality of outlier detection must be evaluated for each application, and the optimal choice might be different depending on the actual spectroscopic technique.

Finally, it should be mentioned that the *i*PLS method was preliminarily tested on an NIR data set of pectins with different degrees of esterification. With 50 *i*PLS intervals, one interval with two PLS components improved the performance of a four-PLS-component full-spectrum model by a factor of 2 with respect to RMSECV.[14] All other intervals gave higher RMSECV values compared to the RMSECV of the full-spectrum model.

CONCLUSION

It might be very useful to select variables or intervals of variables from spectroscopic data ensembles. In this paper a new graphically oriented local modeling approach (*i*PLS) is described and compared to three different variable selection methods by evaluation on a near-infrared spectroscopic data set. For these data it has been shown that *i*PLS is an attractive method in providing an overview of interesting spectral areas which could be selected. The results from using *i*PLS are comparable to the other effective methods tested, but the main contribution from using *i*PLS is the graphic output giving an overview of the spectral data. For specific selection of variables, forward stepwise selection proved to be a good alternative, while methods such as recursively weighted regression and principal variables work well in some cases, depending on the preprocessing of the data. Basically, *i*PLS has proven to represent a sound compromise between data reduction and spectral localization and yet being able to utilize the first-order advantage.

Further research on *i*PLS might include an investigation of all possible combinations of the selected intervals in order to investigate the synergy between different spectral regions. If the number of intervals chosen is less than approximately 20–30, it is possible to evaluate all possible interval combinations depending on how much computer time one can spend. Work is also currently underway with respect to improving and generalizing the interval optimization.[15]

ACKNOWLEDGMENTS

Agnar Höskuldsson, Rasmus Bro, and Carsten Ridder are acknowledged for inspiring discussions on the topic treated in this paper. The investigation is sponsored by Centre for Advanced Food Studies (LMC) with funds for Søren Balling Engelsen and by funds to Professor Lars Munck from the Danish Research Councils 13-4804-1 (agriculture) and 16-5180-1 (technology) and Centre for Predictive Multivariate Process

Analysis. Carlsberg are thanked for making available the analyzed beers including determination of extract concentration.

1. H. Martens and S. A. Jensen, "Partial Least Squares Regression: A New Two Stage NIR Calibration Method", in *Progress in Cereal Chemistry and Technology,* J. Holas and J. Kratochvil, Eds. (Elsevier, Amsterdam, 1983), pp. 607–647.
2. K. H. Norris, *Near Infrared Technology in Agricultural and Food Industries* (American Association of Cereal Chemists, St. Paul, Minnesota, 1987).
3. L. Munck, L. Nørgaard, S. B. Engelsen, R. Bro, and C. A. Andersson, Chemom. Intell. Lab. Syst. **44,** 31 (1998).
4. S. B. Engelsen and L. Nørgaard, Carbohydrate Polym. **30,** 9 (1996).
5. S. B. Engelsen, JAOCS **74,** 1495 (1997).
6. A. Höskuldsson, Chemom. Intell. Lab. Syst. **23,** 1 (1994).
7. M. Andersson, University of Lund, Sweden, e-mail: Martin.Andersson@teknlk.LTH.se, personal communication.
8. J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Harbor, Michigan, 1975).
9. C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. Li Yue, and G. L. Coté, Anal. Chem. **70,** 35 (1998).
10. G. P. McCabe, Technometrics **26,** 137 (1984).
11. W. J. Krzanowski, Appl. Statist. **36,** 22 (1987).
12. H. Martens and T. Næs, *Multivariate Calibration* (Wiley, New York, 1993), 2nd ed.
13. S. Wold, Technometrics **20,** 397 (1978).
14. S. B. Engelsen, E. Mikkelsen, and L. Munck, Progr. Colloid Polym. Sci. **108,** 166 (1998).
15. C. A. Andersson, "Optimization Approaches to Selection of Ranges of Variables in Bi- and Multi-linear Calibration", in preparation for J. Chemom.

Appendix 2

Light scattering and light absorbance separated by Extended Multiplicative Signal Correction (EMSC). Application to NIT analysis of powder mixtures

Harald Martens, Jesper Pram Nielsen and Søren Balling Engelsen

Abstract

The Extended Multiplicative Signal Correction (EMSC) pre-processing method allows a separation of physical light scattering effects from chemical (vibrational) light absorbance effects in spectra from e.g. powders or turbid solutions. It is here applied to diffuse Near Infrared Transmission (NIT) spectra of mixtures of wheat gluten (protein) and starch (carbohydrate) powders, linearised by conventional log(1/T). Without any correction for uncontrolled light scattering variation between the powder samples, these absorbance spectra could give reasonable predictions of the analyte [gluten], but only when using multivariate calibration with a much more complex model than expected. Standard MSC pre-processing did not work for these data at all; it removed too much analyte information. However, the EMSC pre-processing yielded powder spectra that obeyed Beer's Law more or less as if they had been obtained from transparent liquid solutions, apparently by isolating the chemical light absorption from additive, multiplicative and wavelength-dependent effects of uncontrolled light scattering variations. The model-based EMSC and its converse, the Extended Inverted Signal Correction (EISC), gave rather complete description of the diffuse absorbance spectra, and virtually indistinguishable performance in the calibration set and the test set of samples.

Introduction

Pre-processing for multivariate calibration. Multivariable electromagnetic spectrophotometry in the near or mid-infrared region offers great practical and economical advantages for analysis of large sample series, as demonstrated by diffuse reflectance or transmittance spectroscopy in areas such as agriculture, food technology, pharmaceutics and petrochemistry. Today, such high-speed instruments are routinely designed to yield precise quantitative determination for a variety of chemical and physical properties, using multivariate calibration to solve the selectivity problems caused by the lack of sample preparation and for automatic detection of outliers[1]. Pre-processing of the spectral measurements is used for optimising the subsequent multivariate calibration. An example of this is the common linearization of transmittance T into absorbance log(1/T), which under ideal conditions is linearly related to chemical composition according to Beer's law.

When analysing more or less intact complex samples by diffuse reflectance or transmittance spectroscopy, uncontrolled variations in light scattering is often a dominating artefact, which complicates subsequent quantitative chemical analysis. This undesired scattering variation is due to uncontrolled physical variations in the measured samples – particle size and shape, sample packing, sample surface, etc. If the light scattering could be modelled and corrected for mathematically in a more elaborate pre-processing stage, these problems should be reduced or eliminated. The cost of NIR analysis could then be reduced, because the need for controlled sample preparation could be further reduced, the number of calibration samples could be reduced and the statistical calibration modelling process could be simplified. Moreover, a pragmatic, but reasonably accurate model-based light scatter correction may shed new light on the light scattering processes themselves. If successful, such a method for light scatter correction might also be used for other types of instruments, for example, for reducing the need to remove uncontrolled turbidity prior to UV or VIS spectroscopy in general. The problem is how to describe light scattering mathematically in practice.

Additive and multiplicative models for light scattering. In some simple systems, a purely multiplicative effect of light scattering may be observed. In transmittance spectroscopy of e.g. transparent solutions, a change in the optical path length (e.g. cuvette width) scales the whole absorbance spectrum by a given factor, according to the Beer-Lambert law. In diffuse reflectance of powders under ideal conditions, a variation in the overall light scatter coefficient between the samples likewise scales their chemical light absorption spectra, according to the Kubelka-Munk theory[2]. In either case, if the scaling factor for each sample is known or can be measured, the multiplicative interference effect is easy to correct for by a simple re-scaling.

In other systems, a purely additive effect of light scattering may under certain rare conditions be observed. In visible-range “transflection” spectroscopy in turbid aqueous solutions, a variation in turbidity level sometimes causes a simple baseline shift[3], which may be corrected for by a baseline subtraction, according to Beer’s law.

However, in practical diffuse spectroscopy of complex samples under realistic measurement conditions, such a fixed, purely multiplicative or purely additive modelling of light scattering appears to be an oversimplification. A more extensive modelling of light scattering variations is generally required. This modelling may be done as an implicit part of the multivariate calibration process, or in an explicit pre-processing stage.

Implicit scatter correction during multivariate calibration regression. Pragmatic multivariate calibration techniques[1] can to some extent implicitly compensate for unknown scatter variations. Multivariate techniques based on additive regression models, such as bilinear regression using Partial Least Squares Regression[4], can automatically pick up and account for various unknown types of “scatter” by introducing extra regression components. The price for this implicit scatter compensation is a higher complexity of the calibration models, which then become prone to noise and difficult to interpret, and require more calibration samples[5]. Bilinear additive modelling can be regarded as an implicit Taylor expansion of the underlying, unknown relationships in the data[6], and for multiplicative scatter-affected relationships the first-order (additive) approximation of bilinear models like PCA or PLSR is simply not good enough[1].

Explicit data analytical pre-processing prior to calibration can sometimes eliminate non-relevant systematic sources of covariance and may then lead to more simple and robust regression models. Several additive pre-processing methods exist for removing irrelevant spectral contributions, using covariance-based linear (additive) methods to subtract or down-weigh spectral components expected to represent interferants such as Spectral Interference Subtraction[7], Direct Orthogonalization[8], Orthogonal Signal Correction (OSC)[8;9] and GLS Pre-processing[10]. However, light scatter effects tend to give more or less multiplicative contributions to the spectral data, for which these purely additive pre-processing methods[1] cannot be expected to work very well.

Explicit methods for scatter correction in "dirty" systems. A common pre-processing alternative is to reduce the scattering problems by replacing the input spectra by their first or second derivatives, which will remove between-sample variations in baseline offset and linear baseline trends in the spectra. The trade-off is usually noisier spectra, due to the numerical calculation of the derivatives and the derivative spectra may be more difficult to interpret.

Explicit multiplicative AND additive pre-processing methods have also been developed, e.g. Multiplicative Signal Correction (MSC)[11;12], Piecewise Multiplicative Signal Correction (PMSC)[13], Standard Normal Variate (SNV)[14] and the path length correction method PLC-MS[15]. Particularly the spectral derivative and MSC methods are now being used successfully in many applications, possibly because they have been built into several standard software systems for multivariate calibration. The mentioned data transformations have in common that they are relatively simple and that they can be applied without *a priori* knowledge about the samples and their spectra.

Using prior knowledge in the pre-processing. If spectral regions are present where the target analyte or certain chemical interferants exhibit strong absorption, then the MSC parameter estimation may confuse chemical absorption and physical light scattering effects, with dramatically bad results. An example of this will be demonstrated in the present paper. However, if *a priori* theory or quantitative data exist on the kind of samples and spectra involved, this can enhance the performance of the pre-processing. The simplest way is to use prior spectroscopic knowledge about the constituents' spectra, to ignore (down-weigh) spectral regions where dominating chemical constituents absorb very strongly, when determining how to scatter-correct each spectrum in the MSC.

A more ambitious way to use prior knowledge is to extend the MSC model to include new parameters to account for the physical and chemical phenomena that affect the measured absorbance spectra. One such approach is the *Extended Multiplicative Signal Correction* (EMSC). The basic version of EMSC was originally published by Martens and Stark[7]. In the present study the EMSC is modified and applied to a different kind of spectral data, from powder mixtures.

The inverse of the MSC model was first described by Helland *et al.* who named it Inverted Scatter Correction (ISC)[16]. In preparation for the present study, the method was extended in analogy to the EMSC, and is here referred to as *Extended Inverted Signal Correction* (EISC). In a parallel application study[17] a simpler version of the EISC (ISC with wavelength-dependent scattering correction) was then found to improve the calibration to protein content in wheat kernels from single seed NIT spectra.

The aim of this study is to first describe the EMSC method and compare it theoretically to the MSC, then to test the performance of MSC and EMSC on a set of diffuse transmittance spectra of powder mixtures and to compare the EMSC results to those from EISC.

Theory

Modified Beer-Lambert's law and the EMSC model

For transparent solutions of a set of J absorbing chemical constituents, when Beer's law is obeyed, the *theoretical chemical absorbance spectrum* for sample i over a certain range of wavelength channels $\lambda=1,2,\ldots, \Lambda$, row vector $\mathbf{z}_{\mathrm{i,chem}} =[z_{\mathrm{i,chem},\lambda}, \lambda=1,2,\ldots, \Lambda]$ may be assumed to be a linear combination of the absorbance contributions of the J constituents:

$$\mathbf{z}_{\mathrm{i,chem}} = c_{\mathrm{i1}}\mathbf{k}_1{'} + \ldots + c_{\mathrm{ij}}\mathbf{k}_{\mathrm{j}}{'} + \ldots + c_{\mathrm{iJ}}\mathbf{k}_{\mathrm{J}}{'} \qquad \text{(eq. 1)}$$

where c_{ij} is the concentration and column vector $\mathbf{k}_{\mathrm{j}} =[k_{\lambda,\mathrm{j}}, \lambda=1,2,\ldots, \Lambda]'$ the absorptivity spectrum of the jth constituent. Under near-ideal conditions, with fixed optical path length, the *measured absorbance spectrum* for sample i, row vector $\mathbf{z}_{\mathrm{i}}=[z_{\mathrm{i},\lambda}, \lambda=1,2,\ldots, \Lambda]$, is $\mathbf{z}_{\mathrm{i}} \approx \mathbf{z}_{\mathrm{i,chem}}$.

If the constituent spectra $\mathbf{k}_{\mathrm{j}}$, $j=1,2,\ldots,J$ are sufficiently distinct to be linearly independent of each other, eq. (1) may be used for quantitative analysis in multicomponent systems based on multivariate calibration[1], assuming that the data have an additive chemical information structure.

To approximate physical effects related to light scatter variations, the measured absorbance spectra $\mathbf{z}_i$ of each sample *i* may be modelled as a scaled version of the ideal spectrum $\mathbf{z}_{i,chem}$ according to Lambert´s law, or some other *linear transformation*. Moreover, the light scattering effect depends on the wavelength $\boldsymbol{\lambda}$, for which reason a smooth, polynomial wavelength dependency should also be taken into account.

The EMSC model may be written:

$$\mathbf{z}_i \approx a_i + b_i\mathbf{z}_{i,chem} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 \qquad \text{(eq. 2)}$$

where the coefficients a_i and b_i represent the baseline offset a_i and the path length b_i, relative to the baseline offset and path length in a *reference spectrum*. Coefficients d_i and e_i allow for unknown, smoothly wavelength–dependent spectral variations from sample to sample.

If the coefficients in eq. (2) had been known theoretically, or estimated perfectly, then the EMSC correction

$$\mathbf{z}_{i,corrected} = (\mathbf{z}_i - a_i - d_i\boldsymbol{\lambda} - e_i\boldsymbol{\lambda}^2)/b_i \qquad \text{(eq. 3)}$$

would remove the baseline and path length variations as well as the wavelength–dependent spectral effects, yielding corrected spectra with only chemical absorbance information left; $\mathbf{z}_{i,corrected} \approx \mathbf{z}_{i,chem}$. Ideally, it would then be advantageous to replace the measured spectra $\mathbf{z}_i$ with $\mathbf{z}_{i,corrected}$ in subsequent multivariate calibration, since the latter have a simpler and more linear relationship to the analyte concentration. Unfortunately, the parameters are usually unknown and have to be estimated from the available spectrum $\mathbf{z}_i$.

EMSC parameter estimation

The success of EMSC requires that good statistical *estimates* of the model parameters a_i, b_i, d_i and e_i in eq. 2) are found from the measured spectrum $\mathbf{z}_i$ in such a way that they are insensitive to variations in the unknown chemical constituents' concentrations c_{i1}, c_{i2}, ..., c_{iJ}. This is done by including a quantitative description of the constituents' spectra $\mathbf{k}_1$, $\mathbf{k}_2$,..., $\mathbf{k}_J$ in the EMSC model.

If the model for a sample's ideal chemical absorbance spectrum $\mathbf{z}_{i,chem}$ (eq. 1) is inserted directly into the *physical* model (eq. 2), then the model obtains non-linear parameter products $b_ic_{i1}, b_ic_{i2},.., b_ic_{iJ}$. This complicates the parameter estimation, because the important path length parameter b_i cannot be observed and estimated independently of the unknown chemical compositions $c_{i1}, c_{i2},.., c_{iJ}$. In order to have only a sum of linear parameters, Beer's Law expression in eq. (1) can instead be rewritten in terms of variations around a chosen reference spectrum, which will be termed row vector **m**:

$$\mathbf{z}_{i,chem} = \mathbf{m} + \Delta c_{i1}\mathbf{k}_1' + \Delta c_{i2}\mathbf{k}_2' + ... + \Delta c_{iJ}\mathbf{k}_J' \qquad \text{(eq. 4a)}$$

where **m** is some reference spectrum, for example, measured in a "typical" sample or computed as the mean of a set of spectra, and Δc_{ij} represents the difference in constituent # *j*'s concentration between the sample *i* and reference **m**. Equation (4a) may be inserted into the physical model (eq. 2), which now gets a purely additive term $b_i\mathbf{m}$. This construction removes the problem of parameter products b_ic_{i1} $b_ic_{i2},.., b_ic_{iJ}$ as b_i is obtained without other unknown terms. However this generates another, more statistical problem of "collinearity": If the reference spectrum **m** in eq. (4a) is more or less equal to a linear combination of the constituent spectra $\mathbf{k}_1, \mathbf{k}_2,...,\mathbf{k}_J$, then it will be impossible for the parameter estimation to distinguish clearly between the contributions b_i from the reference **m** and the combined contributions Δc_{ij} from all the individual constituents, even if the constituent spectra themselves, $\mathbf{k}_j, j=1,2,...,J$, are linearly independent of each other.

This collinearity problem between reference **m** and the constituent spectra $[\mathbf{k}_1, \mathbf{k}_2,...,\mathbf{k}_J]$ may be overcome in different ways[7], for example by replacing the *J* constituent spectra with *J*-1 (or fewer) linear combinations of them. This new set of "chemical constituent spectra" to be used in the EMSC model may be obtained by singular value decomposition of the centred constituent matrix $\mathbf{G}=[(\mathbf{k}_1 - \mathbf{m}), (\mathbf{k}_2 - \mathbf{m}), ..., (\mathbf{k}_J - \mathbf{m})]$, using only the loadings/eigenvectors that correspond to clearly non-zero singular values.

The binary case: Assume, for example, that the samples are expected to contain only two main chemical constituents, with different, linearly independent spectra $\mathbf{k}_1$ and $\mathbf{k}_2$. Then, if **m** represents their average spectrum, there is only one non-zero singular value in $\mathbf{G}=[(\mathbf{k}_1 - \mathbf{m}), (\mathbf{k}_2 - \mathbf{m})]$. Its loading is proportional to $\mathbf{k}= \mathbf{k}_1 - \mathbf{k}_2$. Thus eq. 1) may be rewritten for this two-constituent model, as

$$\mathbf{z}_{i,chem} = \mathbf{m} + \Delta c_i\mathbf{k} \qquad \text{(eq. 4b)}$$

In the example of gluten/starch mixtures studied in this paper, the pure constituent spectra were defined from two calibration samples known to represent the pure constituents, namely a spectrum of pure gluten (sample 3) and a sample of pure starch (sample 93). With $\mathbf{k}_1 = \mathbf{k}_{\text{Gluten}} = \mathbf{z}_3$ and $\mathbf{k}_2 = \mathbf{k}_{\text{Starch}} = \mathbf{z}_{93}$, the reference spectrum $\mathbf{m}$ was defined by their average, $\mathbf{m} = (\mathbf{k}_{\text{Gluten}}' + \mathbf{k}_{\text{Starch}}')/2$, and the chemical variation spectrum $\mathbf{k}$ by their difference, $\mathbf{k} = \mathbf{k}_{\text{Gluten}}' - \mathbf{k}_{\text{Starch}}'$.

A quantitative understanding of the resulting EMSC model may in this example be obtained from the fact that the constituent concentrations here add up to 1; $c_{\text{i,Gluten}} + c_{\text{i,Starch}} =1$. With eq. (1) rewritten as $\mathbf{z}_{\text{i,chem}} = c_{\text{i,Gluten}}\mathbf{k}_{\text{Gluten}}' + c_{\text{i,Starch}}\mathbf{k}_{\text{Starch}}'$, then eq. (4b) may be rewritten more explicitly as

$$\mathbf{z}_{\text{i,chem}} = \mathbf{m} + (c_{\text{iGluten}}-0.5)\mathbf{k}' \qquad \text{(eq. 4c)}$$

This binary mixture model in eq. (4b or 4c) may now be inserted into the physical model in eq. (2), yielding a linear statistical model with only additive terms, even for b_i:

$$\mathbf{z}_i = a_i\mathbf{1} + b_i\mathbf{m} + h_i\,\mathbf{k}' + d_i\boldsymbol{\lambda} - e_i\boldsymbol{\lambda}^2 + \boldsymbol{\varepsilon}_i \qquad \text{(eq. 5)}$$

where vector $\mathbf{1}$ =[1,1,1,....,1] is introduced for matrix formality. Vector $\boldsymbol{\varepsilon}_i$ is added to represent the residual spectrum of sample *i*, containing random measurement noise and possible un-modelled spectral structures. Note that $h_i = b_i\,\Delta c_i. = b_i\,(c_{\text{iGluten}}-0.5)$. Vectors $\mathbf{m}$ and $\mathbf{k}'$ were already assumed to be sufficiently different from each other. Ideally, all the five row "spectra" or "model vectors" $\mathbf{1}$, $\mathbf{m}$, $\mathbf{k}'$, $\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}^2$ in eq. 5 should be clearly linearly independent of each other. Then the EMSC parameters in vector $\mathbf{p}_i = [a_i, b_i, h_i, d_i \text{ and } e_i]$ (eq. 5) may be estimated by least squares regression of each input spectrum $\mathbf{z}_i$ to the model regressor matrix $\mathbf{M}=[\mathbf{1}; \mathbf{m}; \mathbf{k}'; \boldsymbol{\lambda}; \boldsymbol{\lambda}^2]$ according to the regression model $\mathbf{z}_i = \mathbf{p}_i\,\mathbf{M} + \boldsymbol{\varepsilon}_i$.

A versatile solution for the EMSC parameters in sample *i* is the conventional weighted least squares estimator

$$\mathbf{p}_i = \mathbf{z}_i\mathbf{V}\mathbf{M}'\,(\mathbf{M}\mathbf{V}\mathbf{M}')^{-1} \qquad \text{(eq. 6)}$$

where the diagonal matrix $\mathbf{V}$ allows different weights for different wavelengths. Subsequently, the fit of the individual spectra to the model may be assessed by summarising the estimated residual spectrum $\boldsymbol{\varepsilon}_i = \mathbf{z}_i - \mathbf{p}_i\,\mathbf{M}$, for each sample i=1,2,... . The weights may be defined on the basis of prior knowledge, or by default set to 1.0.

After the unknown physical and chemical parameters [a_i, b_i, h_i, d_i, e_i] have been estimated for every sample in the calibration and test sets, their corrected spectra $\mathbf{z}_{\mathrm{i,corrected}}$, i=1,2,...,100 may be obtained by eq. (3). These corrected spectra may then be used as $\mathbf{X}$ in subsequent multivariate calibration for $\mathbf{y}$ = [gluten].

If there are more than two chemical constituents, an expression similar to eq. 5 may be obtained by inserting eq. 4a) into eq. 2. The EMSC model will then have more than one chemical difference spectrum, and more than one chemical parameter h_{i1}, h_{i2}, ..., but eq. 6 may still be used for the parameter estimation, and eq. 3 for the scattering correction.

Causality and approximation. If the EMSC model in eq. (5) and EMSC parameter estimation in eq. (6) were perfect, the corrected spectra would be linearly related to the concentration of the chemical analyte, in this case $\mathbf{z}_{\mathrm{i,corrected}} = \mathbf{m} + (c_{\mathrm{i,Gluten}}-0.5)\mathbf{k}' + \boldsymbol{\varepsilon}_\mathrm{i}/b_\mathrm{i}$, and this would be optimal for subsequent linear multivariate calibration modelling. In practice, the EMSC model will usually not be causally perfect, due to, for example, un-modelled constituent interactions, stray light and more intricate wavelength dependencies. Nor will the parameter estimates be perfect, due to for example measurement noise in $\mathbf{m}$, $\mathbf{k}$ and/or $\mathbf{z}_\mathrm{i}$. However, while the modelling inside the EMSC may be rather ambitious, the EMSC correction itself is rather conservative, since eq. (2) also passes un-modelled information $\boldsymbol{\varepsilon}_\mathrm{i}$ from $\mathbf{z}_\mathrm{i}$ to $\mathbf{z}_{\mathrm{i,corrected}}$ (although scaled by factor b_i). As a consequence, the EMSC correction (eq. 3) may be seen as just an approximation tool, analogous to the subsequent multivariate calibration modelling itself: If the EMSC is found to simplify the structure without losing too much valuable information, then it has had a net positive effect on the calibration process as a whole.

The EMSC is intended primarily as a spectral pre-processing method to simplify the subsequent multivariate calibration, where $\mathbf{z}_{\mathrm{i,corrected}}$ in a set of samples i=1,2,...,N is used as $\mathbf{X}$ in multivariate calibration for some chemical constituent $\mathbf{y}$. However, the EMSC parameters themselves may also provide valuable information. First of all, if the EMSC model gives a sufficiently complete description of the absorbance spectra, a reasonably good estimate of the analyte concentration may be obtained already during the EMSC pre-processing, before any multivariate calibration. In the present case a good EMSC modelling would yield good gluten estimates from $c_{\mathrm{i,Gluten}} = h_\mathrm{i}/b_\mathrm{i} + 0.5$. Secondly, if the desired information in the absorbance measurements $\mathbf{z}_\mathrm{i}$ are of a physical, rather than chemical nature, then the parameter estimates a_i, b_i, d_i and/or e_i may be used for extracting various types of physical variations. For instance, the matrix of EMSC parameters from eq. (6), [$\mathbf{p}_i$; i=1,2,..., N], could be used as **X** in the calibration step to describe a physical property **y**, such as particle size, sample packing, cuvette thickness etc.

Conventional MSC

The conventional MSC may be seen as a simplification of the EMSC. It assumes the same structure as the EMSC in a reduced version of eq. (2),

$$\mathbf{z}_i \approx a_i + b_i\mathbf{z}_{i,chem} \qquad \text{(eq. 7)}$$

However, the MSC uses a much simpler model of $\mathbf{z}_{i,chem}$ than eqs. (4a-c), as it assumes $\mathbf{z}_{i,chem} \approx \mathbf{m}+\boldsymbol{\delta}_i$, where $\boldsymbol{\delta}_i$ represents “unknown and irrelevant types of variation”, including the spectral contributions due to chemical constituents. For estimating the MSC parameters a_i and b_i, $\boldsymbol{\delta}_i$ is simply ignored and the basic MSC model is therefore written

$$\mathbf{z}_i = a_i\mathbf{1} + b_i\mathbf{m} + \boldsymbol{\varepsilon}_i \qquad \text{(eq. 8)}$$

With the simplified model regressor matrix **M**=[**1**; **m**] with the only two rows presumably linearly independent, eq. (6) yields estimates of the physical a_i and b_i parameters. The MSC correction, of course, is correspondingly simpler form, and eq. 3 reduces to $\mathbf{z}_{i,corrected} = (\mathbf{z}_i - a_i)/b_i$. The MSC may in many cases give useful light scattering correction, because the unknown chemical contributions $\boldsymbol{\delta}_i$ are often relatively uncorrelated with vectors **1** and **m** and hence do not affect the estimation of parameters a_i and b_i very much. However, major spectral effects due to changes in the chemical composition of the samples may render MSC inappropriate.

The results from the inverse version of EMSC, the EISC, will be briefly summarised at the end of the Results and Discussion section.

Material and Methods

Spectral measurement: Transmittance spectra (T) in the range of 850-1050 nm were collected using an Infratec 1255 Food and Feed Analyzer (FossTecator, Höganäs, Sweden) fitted with a standard sample holder for 5 cylindrical cuvettes. A tungsten lamp (50 W) and a diffraction grating were used to create monochromatic light. The light passed through the powders and reached the silicon detector, and transmittance was recorded as $T=I/I_0$.

Samples: Industrial-grade wheat gluten (approx. 80% protein) and pure wheat starch (Merck Eurolab 73502-250) were used for the binary mixture design. Five mixtures with different ratios of gluten/starch (by weight, Table I) were prepared by weighing on a precision balance, and mixed thoroughly.

Table I. Experimental design for the *N*=100 measured spectra

Samples	Mixture	[gluten]/[starch]	Packing	Reps 1,2	Use
1-10	1	1/0	loose	1-5, 6-10	Calibr.
11-20	1	1/0	firm	11-15, 16-20	Calibr.
21-30	2	0.75/0.25	loose	21-25, 26-30	Test
31-40	2	0.75/0.25	firm	31-35, 36-40	Test
41-50	3	0.5/0.5	loose	41-45, 46-50	Calibr.
51-60	3	0.5/0.5	firm	51-55, 56-60	Calibr.
61-70	4	0.25/0.75	loose	61-65, 66-70	Test
71-80	4	0.25/0.75	firm	71-75, 76-80	Test
81-90	5	0/1	loose	81-85, 86-90	Calibr.
91-100	5	0/1	firm	91-95, 96-100	Calibr.

For each of these mixtures, approximately 2 g was taken out randomly in five sampling replicates, and filled loosely into five different glass cuvettes (horisontal diameter of 25mm) to a vertical sample thickness of about 8mm. The NIT spectra of the five sampling replicates were measured vertically, in two consecutive spectral replicates. Then the powder in each of the five cuvettes was packed more firmly, compressed by hand to a sample thickness of about 5-6 mm, and their NIT spectra measured again in two consecutive spectral replicates. In total, the five mixtures x two powder packings x two spectral replicates x five sample holders/powder sampling replicates amounted to a factorial design with *N*=100 samples (5 x 2 x 2 x 5 =100 spectra).

Data analysis: Each transmittance spectrum T was first changed into absorbance $A = \log_{10}(1/T)$. These absorbance spectra were then subjected to various pre-processing methods. The corrected spectra from eq. (3) were submitted to multivariate calibration and prediction[1]: The 100 wavelength channels between 850 and 1050 nm were used in regressor matrix $\mathbf{X}=[\mathbf{z}_{\text{i,corrected}}, i=1,2,...,N]$, and the concentration of the analyte, [gluten], was used as regressand $\mathbf{y}=[c_{\text{i,Gluten}}, i=1,2,...,N]$. Bilinear modelling by Partial Least Squares Regression[4] (PLSR) was used as a low-rank calibration method. To ensure the validity of the obtained results, only the spectra from three of the five mixtures (mixtures 1, 3 and 5; in total 60 spectra) were used as calibration samples for developing the calibration model ($\mathbf{y}\approx f(\mathbf{X})$). The remaining mixtures (mixtures 2 and 4; in total 40 spectra) were used as an "independent" test set, for predictions $\hat{\mathbf{y}}=f(\mathbf{X})$. The predictive validity for models with *0,1,2,...* latent variables (PCs) in the PLSR model was then estimated in two different ways, as described in Martens and Martens[3]: 1) $\text{RMSEP}_{\text{Test}}$ = the prediction error in the independent test set, and 2) RMSEP_{CV} = the corresponding Cross-Validated Root Mean Square Error of Prediction in the calibration set, using a 3-segment version of cross-validation, keeping all replicates for one of the three calibration mixtures out at a time for independent prediction testing. The pre-processing, multivariate calibration and graphics were performed using software written in MatLab version 6.1 (The Matworks, Inc., Natick, MA, USA).

Results and Discussion

No correction for light scattering

Figure 1 shows the performance or the spectral measurements without any pre-processing, i.e. log(1/T). Figure 1A) illustrates traditional univariate calibration: The wavelength with the best correlation to the analyte, $\mathbf{y}$=[gluten], 994 nm ($\mathbf{x}_{73}$), was used as regressor for $\mathbf{y}$ in the calibration samples, and the resulting model was applied to the 40 samples in the independent test set. The figure shows how selectivity problems render such traditional calibration useless in diffuse spectroscopy of light scattering samples, even at the "best" wavelength, which also agrees with the knowledge gained through experience with NIR spectroscopy over the last decades.

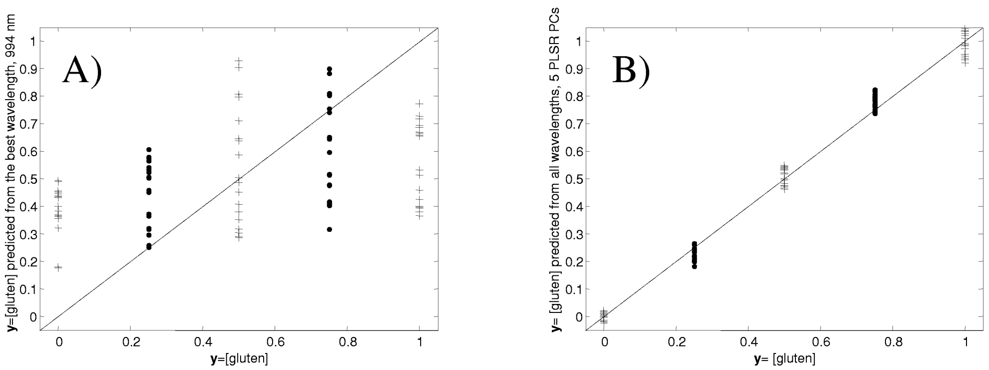


Figure 1. Calibrations without pre-processing of the log(1/T) measurements. A) Univariate calibration, using the best wavelength channel, 994 nm (x_{73}), B) Multivariate calibration, combining all wavelength-channels between 950 and 1050 nm (x_1-x_{100}) via a PLSR model with 5 PCs. Abscissa: Analyte concentration y ([gluten] in powders), ordinate: Prediction at the optimal model rank. Dotted line: Theoretically correct predictions, +=calibration samples, **o**=independent test samples.

In contrast, Figure 1B) shows that multivariate calibration – in this case by PLSR – can overcome most of the selectivity problems, if there are enough calibration samples, and enough PCs are used in the model. The cross-validation within the calibration sample set (+) showed that 5 PCs gave the best compromise between predictive error and model complexity. The figure also shows that this model gives fairly good predictive ability also for the test samples (**o**).

Considering the simplicity of the chemical composition of these gluten/starch mixtures, the complexity of the input spectra shown in figure 2A is surprising. Yet, a 5-PC model (Figure 1B) gave relatively good predictive ability for **y**=[gluten] from these spectra. However, since the samples represent binary mixtures, where the analyte concentrations add up to 1, only one single PC should ideally suffice in the calibration model, if the spectral response had been truly bilinear. Figure 2B clearly illustrates that a single-PC model of the absorbance spectra does not give a satisfactory selectivity for the analyte.

Conventional Multiplicative Signal Correction (MSC)

It is expected that most of the selectivity problems in this study are due to physical variations caused by uncontrolled variations in light scattering due to sample packing, sample surface topology, particle size and possible variations in amount of sample in the cuvettes. In order to separate such physical light scattering variations from the chemical absorbance variations, MSC was first applied to the spectra in Figure 2A. Assuming the simplified chemical model that ignores the

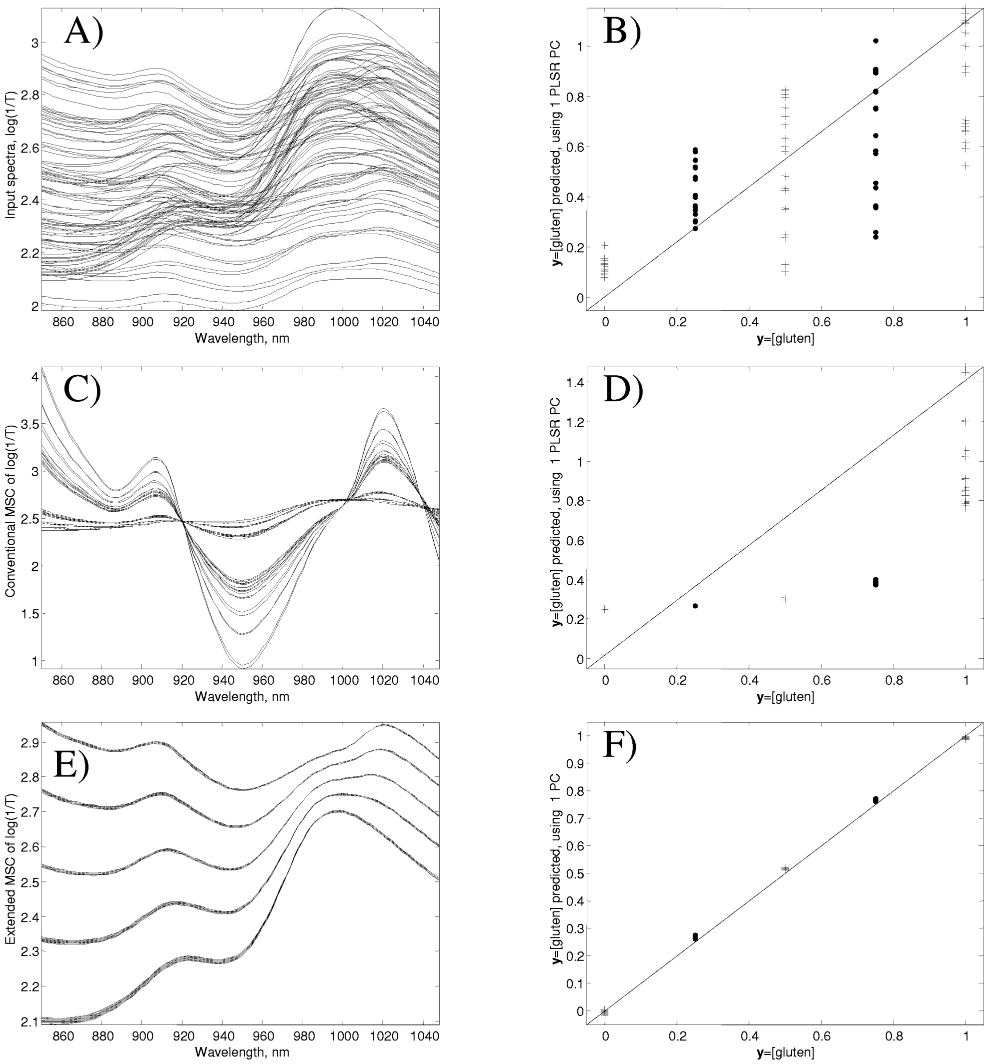


Figure 2. Different pre-processing methods. Left: NIR spectra of 5 gluten/starch mixtures in 20 replicates. Right: Analyte concentration predicted from spectra at the expected model rank (1 PC) vs. input y. Dotted line: Theoretically correct predictions, +=calibration samples, o=independent test samples. A) and B): No pre-processing other than linearization, log(1/T); C) and D): Multiplicative Scatter Correction (MSC) of the log(1/T) spectra; E) and F): Extended Multiplicative Scatter Correction (EMSC) of the log(1/T) spectra.

systematic chemical variations, each calibration and test set spectrum $\mathbf{z}_i$, i=1,2,…,100 was modelled by eq. (8), with parameter estimation according to eq. (6). As usual in MSC, the reference spectrum **m** was defined as the average of the calibration samples' spectra. Changing the weights **V** in eq. (6) did not appreciably improve the present results; hence, for simplicity, equal weights of 1 were used for all wavelength channels. Figure 2C shows the corrected log(1/T) spectra $\mathbf{z}_{i,corrected}$ after MSC pre-processing. While it is clear that the absorbance spectra now look a lot less complex, the variation in some of the spectra (the pure gluten samples) seems to have become exaggerated, while the others have become virtually indistinguishable. This is confirmed in Figure 2D, which shows the predictive performance for **y** from MSC corrected absorbance spectra **X**, using the multivariate calibration model with the expected model complexity, 1 PC. The 20 replicates of the pure gluten samples give very different predictions, while the remaining samples give almost identical and quite erroneous results, both within the calibration set and the test set. Apparently, because the simple MSC ignored the large, systematic chemical variations caused by the gluten and starch differences, $\boldsymbol{\delta}_i$, these differences contaminated the estimates of a_i and b_i in eq. (6), whereby much of the chemical information in the spectra was erroneously removed by the MSC correction.

Extended Multiplicative Signal Correction (EMSC)

Each of the 100 NIT absorbance input spectra log(1/T) in Figure 2A were instead modelled according to eq. (5), submitted to EMSC parameter estimation by eq. (6) and corrected according to eq. (3). The lowest part of Figure 2 shows the effect of the EMSC pre-treatment.The pre-processed spectra are shown in Figure 2E, where the five mixtures are seen as five distinct spectral patterns. The 20 replicate samples for each mixture, so clearly different in the input spectra (Figure 2A), now appear more or less superimposed and indistinguishable.

Figure 2F shows the predictive performance of the EMSC corrected spectra (Figure 2E) using the expected model complexity (1 PC) in the PLSR calibration model. The five mixtures gives gluten concentration predictions close to the theoretical expectation (dotted line). Hence, almost all of the selectivity and linearity problems evident in Figure 2B and 2D have been eliminated by the EMSC pre-processing.

The EMSC pre-treatment in this data set provided a good reduction of the light scattering problems, and this simplified the subsequent calibration for the analyte, **y**=[gluten]. The results are seen to be equally good for the three mixtures used for calibration (+) and the two intermediate mixtures used for the independent test set (**o**). A small response curvature may be observed, indicating room for possible improvement of the EMSC pre-processing or the subsequent calibration.

The problem of the MSC: mixed chemical and physical variation

Figure 3 illustrates the reason for the poor performance of the simple MSC. While the upper half of Figure 3 shows how the spectra behave when all the input spectra resemble each other chemically, the lower half demonstrates how the spectra behave when chemical variations are dominant. Figure 3A shows the twenty replicate input spectra of mixture #3 (containing equal amounts of gluten and starch). The measured samples only differ in physical properties like light scattering. Since the mean $\mathbf{m}_{50/50}$ of the spectra in Figure 3A also represents this 50/50 mixture, the "unknown spectral variability" $\boldsymbol{\delta}_i$ is actually equal to zero, and every spectrum in Figure 3a) should therefore follow the MSC model (eq. 8), $\mathbf{z}_i \approx a_i\mathbf{1} + b_i\mathbf{m}_{50/50}$. Therefore, when one of the spectra $\mathbf{z}_i$ in Figure 3A is plotted against this mean spectrum $\mathbf{m}_{50/50}$, the consecutive wavelength channels should form a series of spots which generates a straight line with a certain offset a_i and a certain slope b_i. Figure 3B shows two spectra selected from Figure 3A (# 41 and 58) together with this mean spectrum $\mathbf{m}_{50/50}$. When the individual spectra are plottet against the mean spectrum (Figure 3C), they are indeed seen to form nice straight lines, from which their different offsets a_i and slopes b_i may be unambiguously estimated.

In contrast, the lower part of the figure illustrates what happens when the input spectra exhibit clear chemical variations (Figure 3D). In this case, two arbitrarily chosen spectra (#3 and 93) (Figure 3E) now display strange wiggles when plotted against the mean spectrum of all the calibration samples **m** (Figure 3F). Clearly, in this case it is difficult to estimate and correct for by spectral slopes b_i and offsets a_i without removing chemical information. The chemical and the physical effects are seriously entangled in these spectra, and it is difficult to isolate wavelength ranges where they can be disentangled. The purpose of the EMSC is to reduce this problem.

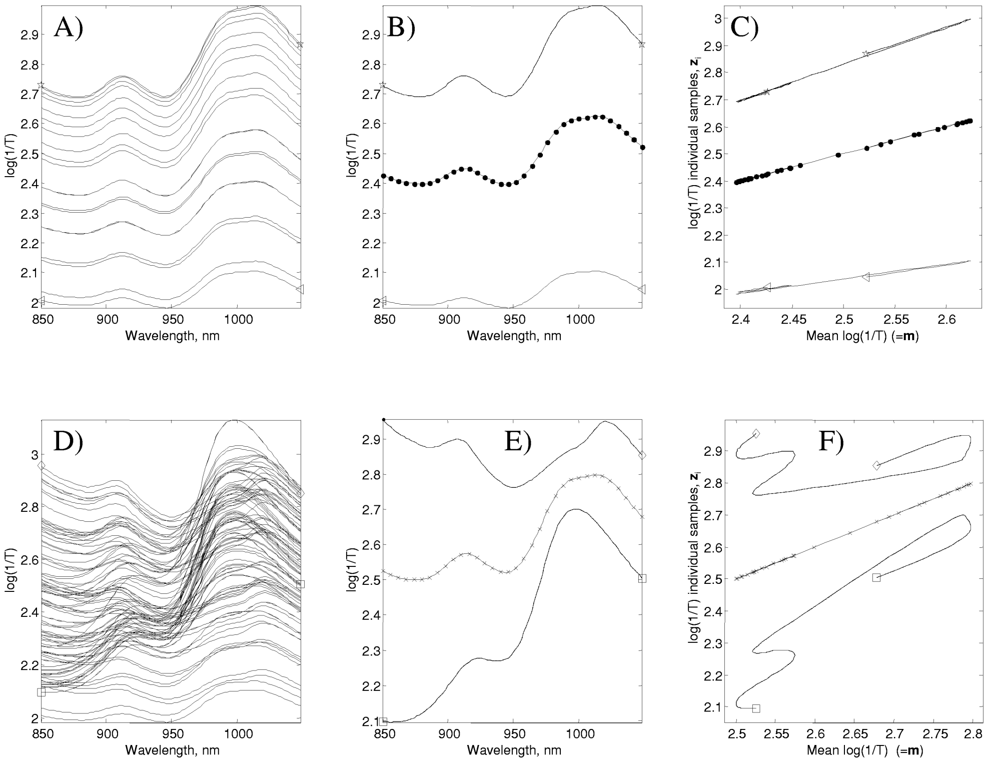


Figure 3. The MSC and its problem when the chemical differences are large. A) NIR spectra of 20 samples of identical chemical composition (50/50 gluten/starch mixture). B) NIR spectra of two of the samples, z_{41} (star) and z_{58} (triangle), and the mean spectrum $m_{50/50}$ of 20 samples (x). C) z_{41} (star), z_{58} (triangle) and $m_{50/50}$ (diagonal x) plotted against $m_{50/50}$ (abscissa). Lower row: Different chemistry, different physics: D) NIR spectra of 20 replicate spectra of each of the 5 mixtures of gluten and starch. E) Two of the NIR spectra, z_3 (diamonds) and z_{93} (squares), and the mean spectrum m of all the spectra (x). F) z_3 (diamonds), z_{93} (squares) and m (diagonal, x) plotted against m (abscissa).

The EMSC model and its parameter estimates

While the EMSC may be used in a software module for ”black box” pre-processing, it is interesting to study the reasons for the good EMSC modelling reported in Figures 2E and 2F. An overview of the EMSC process is given by Figure 4 which shows the spectra in the EMSC model (eq. 5) and the corresponding parameter estimates for the 100 samples obtained by eq. (6). The first row in Figure 4(ACE) illustrates the offset and slope (the “MSC part”) of the

EMSC model. Figure 4A shows vectors **1** and **m** $=((\mathbf{k}_{\mathrm{Gluten}}' + \mathbf{k}_{\mathrm{Starch}}')/2)$, and Figure 4B the corresponding additive and multiplicative parameter estimates a_{i} and b_{i}. These parameter estimates are seen to vary in a complex pattern around their expected values (0 and 1, respectively). For example it is observed that among the first 20 samples (pure gluten), the first 10 (loosely packed) differ markedly from the next 10 (densely packed), illustrating that the EMSC parameters may be used in their own right for quantifying physical properties such as sample packing. This pattern is also seen clearly in the next samples 21-40 and 41-60 representing the 75% and 50% gluten, but less clearly for samples 61-80 and 81-100, i.e. 25% and 0% gluten. Parameter estimates a_{i} and b_{i} are seen to be strongly negatively correlated; $r(a_{\mathrm{i}},b_{\mathrm{i}}) = -0.998$ over all 100 samples. The reason for this correlation is presently unclear, but could reflect instrument geometry; work is in progress to check this.

The second row in Figure 4(BDF) illustrates the chemical part of the EMSC model. The two dotted curves in Figure 4C represent the two input spectra $\mathbf{z}_3$ ($\mathbf{k}_{\mathrm{Gluten}}'$, top) and $\mathbf{z}_{93}$ ($\mathbf{k}_{\mathrm{Starch}}'$) used for defining **m** and the chemical difference spectrum $\mathbf{k} = \mathbf{k}_{\mathrm{Gluten}}' - \mathbf{k}_{\mathrm{Starch}}'$ (solid curve). Figure 4D shows the resulting estimate of the analyte concentration c_{i}, obtained already during the EMSC pre-processing. For comparison, the true analyte fractions, elements y_{i} in vector **y**, are shown as dotted lines. The figure reveals that good concentration estimates c_{i} in this case were obtained already during the EMSC pre-processing, both for the calibration samples (+) as well as for the test samples (**o**).

The third row in Figure 4 shows the wavelength correction part of the EMSC model. Figure 4E shows wavelength vector $\boldsymbol{\lambda}$ as a linear function of the number of nanometers, ranging from –1 to +1, as well as its square. Figure 4F shows their coefficient estimates d_{i} and e_{i}. Both effects are seen to be centred around 0, and their variations are rather small compared to that of offset a_{i} (Figure 4B). A negative correlation between d_{i} and e_{i} is observed; $r(d_{\mathrm{i}},e_{\mathrm{i}})= -0.80$ over all 100 samples, but the reason for this is not yet clear.

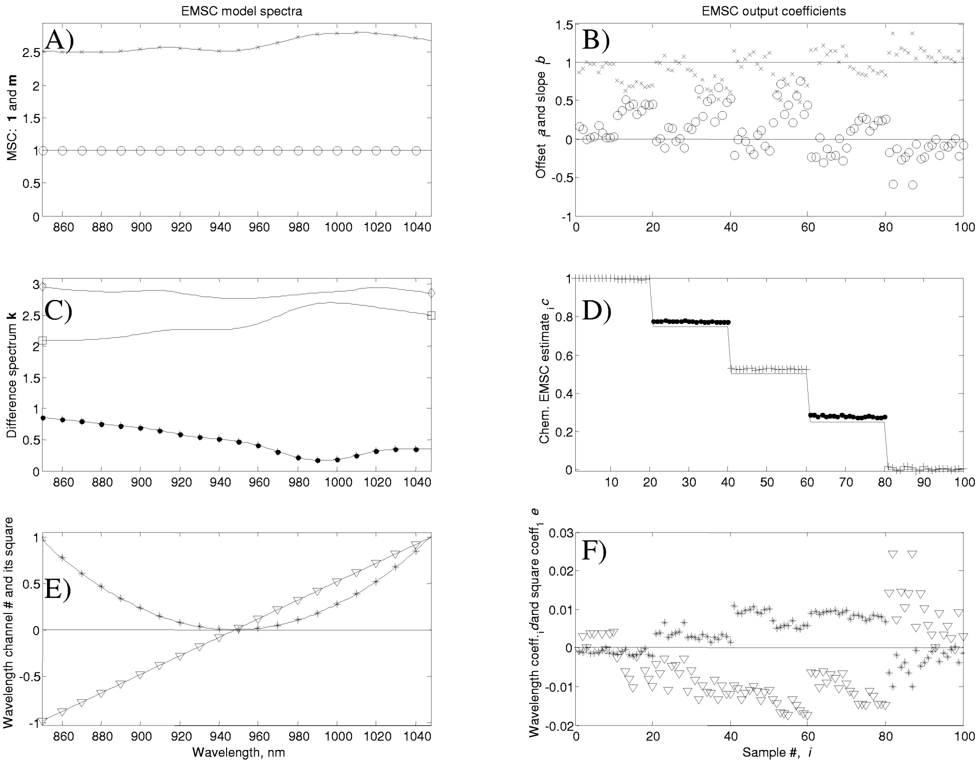


Figure 4. The EMSC model and its estimated parameters. Left: EMSC model spectra (eq.5) as functions of wavelength. Right: Estimated EMSC model parameters from eq. (6) shown as functions of sample #. A) The MSC part of the model: Vector 1 (+++) and reference spectrum m (xxx). B) Additive parameter a_i (+++) and multiplicative parameter b_i (xxx). C) The chemical model extension: Spectra of two sample's, z_3 (diamond) and z_{93} (squares), and the EMSC model vector $k=z_3-z_{93}$ (+). D) Chemical EMSC parameter: Analyte concentration estimated already during the EMSC, c_i, for the calibration samples (*) and the independent test samples ( ). The true analyte fractions are given by the dotted line. E) The physical model extensions: Wavelength index λ (triangles, between –1 and 1) and its square λ^2(asterisk). F) Physical EMSC parameters: Wavelength coefficients e_i (triangle) and wavelength-squared coefficient f_i (asterisk).

How the different pre-processing methods perform for multivariate calibration

Figure 5 summarises the three ways to pre-process the NIT absorbance spectra in terms of their ability to predict **y**=[gluten] when used as **X** in the subsequent multivariate linear calibration modelling by PLSR. Each curve represents a Root Mean Square Error of Prediction (RMSEP), i.e. an “average” of the prediction error in **y,** plotted against calibration model complexity (0, 1, 2,..., 6 PCs in the PLSR model). The three line-symbols (dotted, dashed and solid) were obtained using the three different pre-processing methods reported above: Dotted curves represent the untreated NIT absorbance spectra shown in Figure 2A, dashed curves represent the MSC pre-treated spectra (Figure 2C) and solid curves the new EMSC pre-treated spectra (Figure 2E). Curves marked with the symbol "+" (RMSEP_{CV}) were estimated by the cross-validation within the calibration set, while the curves marked with symbol "o" were obtained for the test set ($\text{RMSEP}_{\text{Test}}$). At 1 PC, the curves summarise the predictions already shown in Figures 2B, 2D, and 2F, respectively.

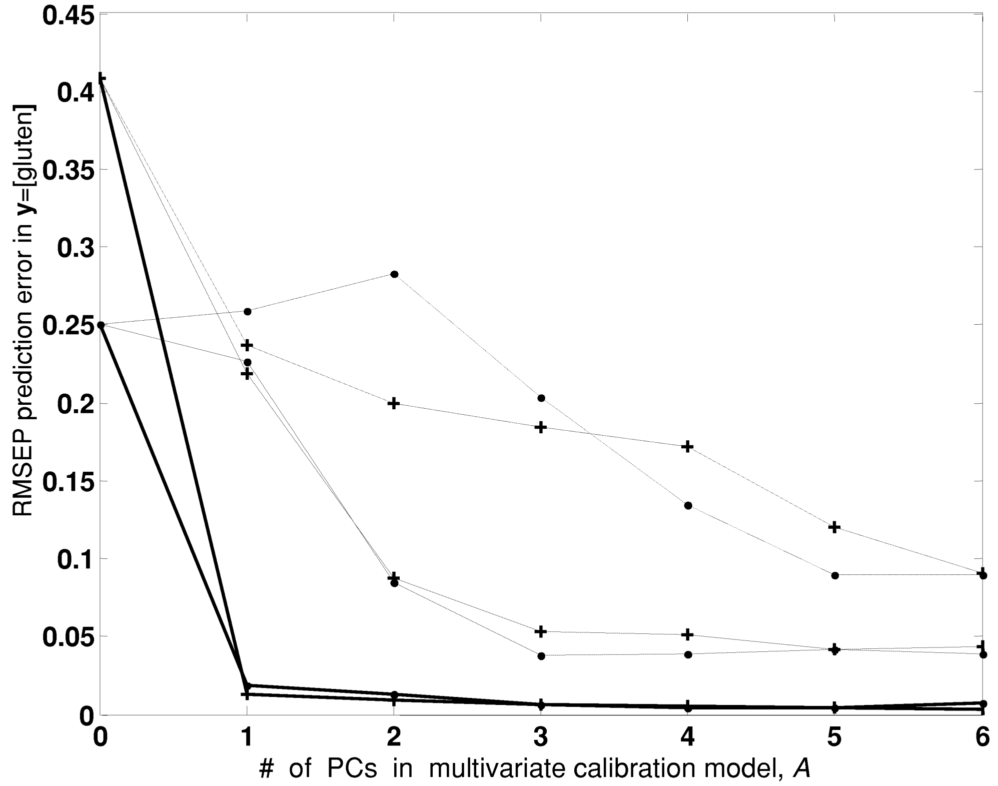


Figure 5. Comparison of pre-processings. Prediction error (RMSEP) for analyte y =[gluten] vs # of PCs in the multivariate calibration. + = Calibration set (mixtures 1, 3 and 5); cross-validation between mixtures. o = Independent test set (mixtures 2 and 4). Pre-processing: … = input absorbance spectra log(1/T), ---- = after MSC, ___ = after EMSC.

As usual, the cross-validation and the test-set methods gave rather similar RMSEP error estimates, when enough PCs were used to avoid underfitting. An example of this is seen at the "optimal" 5 PC solution for the EMSC pre-processed spectra, where RMSEP_{CV} = $\text{RMSEP}_{\text{Test}}$=0.005. Since all RMSEP errors are given in **y**'s concentration scale 0-1, this latter corresponds to an “average” prediction uncertainty of +/-0.5% on a scale of 0-100%, which is quite low, considering the complexity of the original input spectra (Figure 2A).

Figure 5 confirms that the conventional MSC had a detrimental effect in the present case: The subsequent bilinear modelling behaved much worse than the untreated log(1/T) spectra and required more PCs. With the ideal model complexity (1 PC), both of these simple methods displayed poor performance. In contrast, already after 1 PC the EMSC solution yielded very low prediction errors ($RMSEP_{CV}$=0.013 and $RMSEP_{Test}$=0.019). In fact, this solution was better than any of the other solutions using raw log(1/T) or MSC pre-treated spectra. The EMSC model (eq. 5) fitted the 100 spectra quite well. Averaged over all N=100 samples and all K=100 wavelength channels, the variability of the spectra was reduced from a total initial standard deviation of 0.2 absorbance units in the spectra $\mathbf{z}_i$ (Figure 2A) to a total standard deviation of only 0.0005 absorbance units in the EMSC residuals $\boldsymbol{\varepsilon}_i$ (eq. 5).

The choice of samples 3 and 93 to represent the pure constituents in the EMSC model was somewhat arbitrary. Work is in progress to study the effect of using other pairs of samples and to simplify and optimise the EMSC modelling.

Extended Inverse Scatter Correction: EISC

As mentioned previously the EISC is an extension of the ISC method, which switches the roles of $\mathbf{z}_i$ and $\mathbf{m}$ compared to MSC. The ISC model was recently extended with smooth wavelength-dependent terms, $d_i\boldsymbol{\lambda}$ and $e_i\boldsymbol{\lambda}^2$, for successful pre-processing of NIT spectra of intact wheat kernels[17]. Presently, it was further extended with the chemical term $h_i\mathbf{k}$, to yield the Extended ISC (EISC) model with residual vector $\boldsymbol{\gamma}_i$:

$$\mathbf{m} = a_i\mathbf{1} + b_i\mathbf{z}_i + h_i\mathbf{k} + d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2 + \boldsymbol{\gamma}_i \qquad \text{(eq. 9)}$$

After having estimated the parameters a_i, b_i, h_i, d_i and e_i by regressing $\mathbf{m}$ on $\mathbf{M}=[\mathbf{1}\ \mathbf{z}_i\ \mathbf{k}'\ \boldsymbol{\lambda}\ \boldsymbol{\lambda}^2]$ in analogy to eq. (6), the spectra were corrected by $\mathbf{z}_{i,corrected} = a_i\mathbf{1}+b_i\mathbf{z}_i+d_i\boldsymbol{\lambda}+e_i\boldsymbol{\lambda}^2$.

The pre-treated NIT spectra $\mathbf{z}_{i,corrected}$ resulting from the EISC were visually indistinguishable from those of EMSC (Figure 4A), and so was their predictive performance in subsequent multivariate calibration by PLSR (Figures 2B and 5), hence they are not reported here. The similarity between the present EMSC and EISC modelling results are interesting, since the difference between the two methods is analogous to the difference between reverse ("classical") and forward ("inverse") calibration methods, except that EMSC and EISC employ regression over wavelengths, while calibration employs regression over samples. It is well known that forward and reverse calibration[1] give almost identical results when the models fit the data very well. Hence, the small residuals $\boldsymbol{\varepsilon}_i$ from EMSC and $\boldsymbol{\gamma}_i$ from EISC explain the similarity in the results from the EMSC and EISC pre-processing.

In the present study the EMSC (and EISC) method was applied to diffuse transmittance spectra of binary powder mixtures. More work is needed to test the method on more complex data. But this type of pre-processing is expected be useful for diffuse reflectance or transmittance spectra obtained in other spectral ranges, e.g in UV, VIS or IR, and from more complex types of mixtures and for other light scattering materials. Preliminary experience has also shown benefits from applying EMSC or EISC to very different types of data, e.g. in chromatography to correct of uncontrolled variations in baseline offset and total sample concentration, and in descriptive sensory analysis to correct for uncontrolled variations in how individual assessors use the sensory scale (work is in progress in this respect).

In many cases the measured data cannot be expected to be completely modelled already at the pre-processing stage. However, pre-processing could still be quite useful by simplifying the subsequent multivariate calibration regression as well as by revealing something about the nature of the selectivity problems. The ambitious, theory-driven EMSC and EISC pre-processing may then describe and correct for interference phenomena that are more or less expected and understood, while the subsequent data-driven multivariate calibration regression can reveal and correct for unexpected or poorly understood phenomena. In summary, this could be seen as a flexible and powerful combination of deductive and inductive traditions in analytical chemistry.

Conclusion

The EMSC pre-processing simplified a set of diffuse NIT absorbance spectra measured in the lower NIR range by transmittance through 5-8 mm of highly light scattering mixtures of gluten and starch powders. The success is presumably due to the ability of spectral modelling to separate chemical light absorbance and physical light scatter effects. Using prior knowledge about the absorbance spectra of the major constituents, and assumptions about smooth wavelength-dependency of the light scattering variation, the corrected spectra became insensitive to light scattering variations, and responded linearly to the analyte concentration. Thus, the subsequent multivariate calibration regression model became much simpler and had better predictive performance. In fact, the pre-processing proved so effective in the present study that the multivariate calibration regression became superfluous, since the analyte fraction estimate from the EMSC modelling itself provided a direct measure of the desired analyte fraction. Extended Inverted Signal Correction (EISC) yielded corrected spectra and calibration models that were almost indistinguishable from those of the corresponding EMSC.

Acknowledgements

The authors wish to thank Danish Slaughteries and the Ministry of Food, Agriculture and Fisheries for financial support through the project “Early post mortem measurement of WHC (water-holding capacity) and drip loss in fresh pork”. The Danish Cereal Network-B4, and the Advanced Quality Monitoring programme (FELFO) are also acknowledged for financial support. The Matlab programming was supported by Consensus Analysis AS. Gilda Kischinovsky is acknowledged for assistance with the manuscript.

The relationship of EMSC and EISC to the following patents is acknowledged: E. Stark and H. Martens: "Multiplicative Signal Correction Method and Apparatus", US Patent 5,568,400, and E. Stark and H. Martens: "Improved Multiplicative Signal Correction Method and Apparatus" European patent 0415401.

Reference List

1. Martens, H.; Næs, T. *Multivariate Calibration,* Wiley: New York, 1993.
2. Kubelka, P.; Munk, F. *Z.Tech.Phys.* 1931, *12*, 593-604.
3. Martens, H.; Martens, M. *Multivariate Analysis of Quality. An Introduction,* John Wiley and Sons Ltd.: Chichester, UK, 2001.
4. Wold, S.; Martens, H.; Wold, H. *Lecture Notes in Mathematics* 1983, *973*, 286-93.
5. Martens, H.; Næs, T. *Near-Infrared Technology in the Agriculturel and Food Industries*, Williams, P. C.; Norris, K. H., Eds.; 2nd ed.; American Association of Cereal Chemists: St. Paul, 2001; Chapter 4.
6. Wold, S. *Technometrics* 1978, *20*, 397-405.
7. Martens, H.; Stark, E. *Journal of Pharmaceutical and Biomedical Analysis* 1991, *9*, 625-35.
8. Andersson, C. A. *Chemom.Intell.Lab.Syst.* 1999, *47*, 51-63.
9. Wold, S.; Antii, H.; Lindgren, F.; Öhman, J. *Chemom.Intell.Lab.Syst.* 1998, *44*, 229-44.
10. Martens, H.; Høy, M.; Westad, F.; Wise, B.; Bro, R.; Brockhoff, P. B. *J.Chemom.* 2002.
11. Martens, H.; Jensen, S. A.; Geladi, P. *Nordic Symposium on Applied Statistics*, Skagenkaien 12, 1983 6-12-1983; Stokkand Forlag Publ.; 208-34.
12. Geladi, P.; McDougall, D.; Martens, H. *Appl.Spectrosc.* 1985, *39*, 491-500.
13. Isaksson, T.; Kowalski, B. R. *Appl.Spectrosc.* 1993, *47*, 702-09.
14. Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl.Spectrosc.* 1989, *43*, 772-77.
15. Miller, C. E.; Næs, T. *Appl.Spectrosc.* 1990, *44*, 895-98.
16. Helland, I. S.; Næs, T.; Isaksson, T. *Chemom.Intell.Lab.Syst.* 1995, *29*, 233-41.
17. Pedersen, D. K.; Martens, H.; Nielsen, J. P.; Engelsen, S. B. *Appl.Spectrosc.* 2002.

Appendix 3

Near infrared absorption and scattering separated by Extended Inverted Signal Correction (EISC). Analysis of NIT spectra of single wheat seeds

Dorthe Kjær Pedersen, Harald Martens, Jesper Pram Nielsen and Søren Balling Engelsen

Near-Infrared Absorption and Scattering Separated by Extended Inverted Signal Correction (EISC): Analysis of Near-Infrared Transmittance Spectra of Single Wheat Seeds

DORTHE KJÆR PEDERSEN, HARALD MARTENS, JESPER PRAM NIELSEN, and SØREN BALLING ENGELSEN*

Center for Advanced Food Studies, Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

A new extended method for separating, e.g., scattering from absorbance in spectroscopic measurements, extended inverted signal correction (EISC), is presented and compared to multiplicative signal correction (MSC) and existing modifications of this. EISC preprocessing is applied to near-infrared transmittance (NIT) spectra of single wheat kernels with the aim of improving the multivariate calibration for protein content by partial least-squares regression (PLSR). The primary justification of the EISC method is to facilitate removal of spectral artifacts and interferences that are uncorrelated to target analyte concentration. In this study, EISC is applied in a general form, including additive terms, multiplicative terms, wavelength dependency of the light scatter coefficient, and simple polynomial terms. It is compared to conventional MSC and derivative methods for spectral preprocessing. Performance of the EISC was found to be comparable to a more complex dual-transformation model obtained by first calculating the second derivative NIT spectra followed by MSC. The calibration model based on EISC preprocessing performed better than models based on the raw data, second derivatives, MSC, and MSC followed by second derivatives.

Index Headings: **Additive; Multiplicative; Interference; Inverted scatter correction; ISC; Extended inverted signal correction; EISC; Multiplicative signal correction; MSC; Near-infrared transmittance; NIT; Partial least-squares regression; PLSR; Protein; Single seed; Light scattering.**

INTRODUCTION

The extended inverted signal correction (EISC) method, originally developed with chemical analyte extensions,[1] is here presented with spectroscopic extensions. This new method is then applied to near-infrared transmittance (NIT) spectra of single wheat kernels prior to multivariate calibration[2] for protein content, with the calibration model estimated by cross-validated partial least-squares regression (PLSR).[3,4] The most basic version of EISC, ISC, was originally called "inverted scatter correction" (ISC).[5] Martens et al.[1] explains the rationale behind the EISC method and its chemically based extensions, in relation to its heritage, the multiplicative signal correction (MSC)[6,7] and the extended MSC (EMSC).[8] In the present study, the EISC method is extended with some general physical approximation parameters (wavelength dependency and curvatures), and compared to derivative-based preprocessing.

Today, near-infrared (NIR) spectroscopy in combination with multivariate calibration has become the established method for protein determination in cereal breeding as well as for quality determination in the cereal industry, relieving the more than 100-year-old, slow, chemical analysis invented at the Carlsberg Laboratories by the Danish chemist Johan G. Kjeldahl in 1883.[9] The advantages of using NIR spectroscopic methods for cereal quality are mainly the speed of the analysis and their noninvasive character, which is essential if seed fertility is the aim in breeding programs. As an important spin-off, NIR methods provide possibilities for simultaneous determination of additional quality parameters such as moisture, starch, and fiber content.

In low-cost/high-speed analyses of complex systems such as whole-wheat grain, the pattern of optical paths is very complex, and several physical phenomena may contribute to the apparent pattern of "light scatter". The information in NIR spectra usually results from both diffuse light scatter and chemically (vibrationally) absorbed light by the sample, and the NIT spectra of single seeds can be considered a worst case with large additive and multiplicative scatter effects due to differences in kernel size, structure, and presentation angle. It is not uncommon to see more than 95% of the variance in NIR $\log(1/T)$ or $\log(1/R)$ data caused by uncontrolled light scattering variations, which usually will dominate the first latent variable in PCA (principal component analysis) or PLSR modeling. In some cases this is desirable, as when the quality to be calibrated for is physical and related to light scattering, e.g., hardness variation of wheat kernels or particle size variation in powders. However, in most cases light scattering creates selectivity and linearity problems for simple quality attributes related to chemical concentrations. In such cases it is imperative that scatter is isolated from the NIT spectra prior to calibration in order to provide a robust and accurate quantitative method.

The single seed protein system has been studied in depth by Delwiche,[10] who found that an optimal data transformation prior to PLSR calibration was obtained by first calculating the second derivative spectra and then correcting them by MSC. The performance of this double transformation model is confirmed by this study,[11] but such a complex pre-transformation naturally calls for the development of more general and powerful pre-transformations. In the present study it is demonstrated that a general form of EISC is able to provide a quantitative protein model with a precision equal to Delwiche's doubly pre-transformed model.

THEORY

In its most basic "ISC" form, the EISC data transformation can correct a combination of additive and multipli-

Received 13 February 2002; accepted 22 April 2002.

* Author to whom correspondence should be sent.

0003-7028/02/5609-1206$2.00/0

cative interference effects in measured spectra, analogous to the original MSC method.[6,7] Both the MSC and the ISC/EISC adjust the input spectrum of each sample, $\mathbf{z}_i$ in a set of samples, $i = 1, 2, \ldots$, towards a common reference spectrum, $\mathbf{m}$, in order to separate possible physical effects from possible chemical absorption effects. The difference between the methods is that the ISC simply reverses regressor and regressand in each sample's regression model between $\mathbf{z}_i$ and $\mathbf{m}$. Like the conventional MSC, the ISC ("basic EISC") preprocessing method estimates and isolates two presumably physical effects for each sample: an *additive* baseline offset effect and a *multiplicative* scaling effect. If the input information $\mathbf{z}_i$ represents absorbance ($A = \log(I_o/I) = \log(1/T)$) values, the additive baseline offset is intended to model an unknown, fixed amount of absorbance lost at every wavelength, e.g., due to light failing to reach the detector because of dispersion of light in the sample. Multiplicative scaling is intended to model an unknown amplification of the absorbance at every wavelength, e.g., due to a change in the effective optical path length because of light scattering effects in the sample.

In the original EISC paper,[1] the basic version of EISC was extended with *chemical* information known *a priori* to represent absorbance spectra from interfering constituents. In the present paper the EISC is instead extended with *physical* information representing wavelength and polynomial extensions when compared to MSC.

Multiplicative Signal Correction. The multiplicative signal correction, originally named multiplicative scatter correction, MSC, involves correcting each input spectrum $\mathbf{z}_i = [z_{i1}, z_{i2}, \ldots, z_{ik}, \ldots, z_{iK}]$ in a set of related samples $i = 1, 2, \ldots$ towards an ideal spectrum $\mathbf{m}$ where the influence of physical scattering variations has been removed from the effects of chemical absorbance (K is the number of variables in the spectrum). The basic MSC consists of estimating two coefficients, a_i and b_i that ideally contain all the *physical* information in $\mathbf{z}_i$, based on the linear regression model

$$\mathbf{z}_i = a_i + b_i\mathbf{m} + \boldsymbol{\epsilon}_i \tag{1}$$

where $\boldsymbol{\epsilon}_i = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k, \ldots, \varepsilon_K]$ are the residuals that ideally contain all the *chemically* relevant information in $\mathbf{z}_i$, plus other unmodeled effects and random noise. Vector $\mathbf{m} = [m_1, m_2, \ldots, m_k, \ldots, m_K]$ is a common reference spectrum. After parameters a_i and b_i have been estimated, the corrected spectrum $\mathbf{z}_{i,\text{corrected}}$ for this sample is then generated by reversing Eq. 1, from the estimates of a_i and b_i, in an analogy to the univariate "reverse"[2] calibration:

$$\mathbf{z}_{i,\text{corrected}} = (\mathbf{z}_i - a_i)/b_i \tag{2}$$

These corrected spectra may be used as regressors in the subsequent multivariate calibration modeling of the analyte y_i from $\mathbf{x}_i = \mathbf{z}_{i,\text{corrected}}$ over a set of samples, $y_i = f(\mathbf{x}_i)$, $i = 1, 2, \ldots$.

The common reference spectrum $\mathbf{m}$ in Eq. 1 may, for example, be defined as the mean of a set of N spectra of calibration samples:

$$\mathbf{m} = \frac{\sum_{i=1}^{N} \mathbf{z}_i}{N} \tag{3}$$

This reference spectrum $\mathbf{m}$ from the spectra of the N calibration samples $\mathbf{z}_i$, $i = 1, 2, \ldots, N$ may also be applied to MSC of new spectra $\mathbf{z}_i$, $i = 1, 2, 3, \ldots$ e.g., from future prediction samples; Eqs. 1 and 2 are the same for both kinds of samples.

In each calibration or prediction sample *i,* the unknown additive and multiplicative MSC coefficients a_i and b_i in Eq. 1 may be estimated by ordinary least-squares regression of $\mathbf{z}_i$ on $\mathbf{m}$, minimizing the sum of squared residuals in $\boldsymbol{\epsilon}_i$.

$$[a_i \quad b_i] = ([1 \quad \mathbf{m}']'[1 \quad \mathbf{m}'])^{-1}[1 \quad \mathbf{m}']'\mathbf{z}_i' \tag{4}$$

However, the process for estimating and correcting for scattering parameters a_i and b_i is only safe if the effects of chemical variation between $\mathbf{z}_i$ and $\mathbf{m}$ can be ignored; otherwise, the coefficients a_i and b_i may be contaminated with information about, e.g., the analyte, which will then be partially lost in $\mathbf{z}_{i,\text{corrected}}$ (Eq. 2). If applied to pure baseline separated absorbance bands, MSC (and basic EISC) will remove all relevant chemical information, as concentration will have a simple multiplicative effect on the spectral band. For this reason it is good practice to test the scatter coefficients a_i and b_i for information about the analyte or quality to be calibrated for. If a_i and b_i are found to be informative, they may even be included as additional regressor variables in the subsequent multivariate calibration models.

The problem of mixing chemical and physical information in the MSC may alternatively be reduced by down-weighting wavelength regions that carry chemical information. However, in some applications, like NIT of single wheat grains in the 850–1050 nm range, it is difficult to find wavelength regions that are sufficiently informative about the light scattering, but which do not carry chemical information. A more elegant approach would be an MSC model that includes information about the spectra of the chemical constituents.[1,2,8] However, in some applications like the present one, the *in situ* constituent spectra are not known and are difficult to measure.

Basic Extended Inverted Signal Correction. The basic form of the EISC, ISC, is similar to MSC, but it may be more flexible and easier to understand for spectroscopists using multivariate calibration modeling by, for instance, PLSR. MSC, like its extensions, is based on a "reverse"[8] correction in Eq. 2, compared to the model in Eq. 1. In contrast, the basic EISC, like its extensions, uses a "forward"[8] model: the *same* direction of the relationship between spectrum $\mathbf{z}_i$ and reference spectrum $\mathbf{m}$ is kept, both in the model specification:

$$\mathbf{m} = a_i + b_i\mathbf{z}_i + \boldsymbol{\epsilon}_i \tag{5}$$

and in the final correction of the spectra:

$$\mathbf{z}_{i,\text{corrected}} = a_i + b_i\mathbf{z}_i \tag{6}$$

Hence, instead of regressing $\mathbf{z}_i$ on $\mathbf{m}$ in the model (Eq. 1) and then reversing this model in the signal correction step (Eq. 2), the inversed MSC (ISC/EISC) regresses $\mathbf{m}$

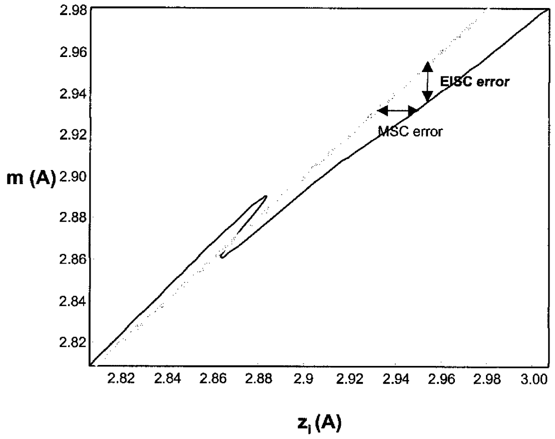


FIG. 1. Plot of a wheat kernel sample's input spectrum $\mathbf{z}_i$ vs. the mean ($\mathbf{m}$) spectrum of the NIT wavelength range (850–1050 nm). While the MSC models the error horizontally on the individual spectra, the EISC models the error vertically on the average spectrum.

on $\mathbf{z}_i$ and uses this "forward" model directly in the signal correction step (Eq. 6).

The estimation of the parameters a_i and b_i may be done by ordinary least-squares regression

$$[a_i \quad b_i] = ([1 \quad \mathbf{z}_i]'[1 \quad \mathbf{z}_i])^{-1}[1 \quad \mathbf{z}_i]'\mathbf{m} \tag{7}$$

Like in MSC, weighted least-squares regression may be used instead if certain wavelengths are to be eliminated because of too-strong overlap between constituent spectra and light scattering effects. The rationale behind this model is explained by Martens et al.[1] The statistical difference between EISC and MSC in their basic form, discussed more theoretically by Helland et al.,[5] is illustrated in Fig. 1. In the plot of reference spectrum $\mathbf{m}$ vs. a sample's input spectrum $\mathbf{z}_i$, the residuals ε_{ik} are minimized horizontally in MSC (noise modeled on the individual spectra) and vertically in ISC/EISC (noise modeled on the average/reference spectrum).

General Spectroscopic Extensions of EISC. Just as the MSC can be extended into extended multiplicative signal correction (EMSC),[8] the basic EISC can be extended to accommodate various types of physical or chemical *a priori* knowledge. In the present case, it is impossible to find wavelength ranges that distinguish the physical light scattering information from the chemical absorbance information. That must be expected to create problems for the MSC or basic EISC methods, but EMSC as well as EISC extended with chemical constituent spectra might solve the problems. However, contrary to the case in Martens et al.,[1] the present *in situ* chemical constituent spectra are not known. In the grain, water is probably bound to a greater or lesser extent to the protein, starch, and cellulose biopolymers, and the NIR *in situ* spectral contributions from the constituents may therefore be rather different from those of isolated constituents in a pure state.

On the other hand, the heterogeneity of the intact wheat grains may cause rather complex *optical* phenomena that are difficult to model explicitly, but which may be approximated in more detail by polynomial extension of Eq. 5, e.g.,

$$\mathbf{m} = a_i + b_i\mathbf{z}_i + c_i\mathbf{z}_i^2 + \boldsymbol{\epsilon}_i \tag{8}$$

Moreover, we expect the light scattering coefficient to have some dependency on the wave number. The exponent of this dependency depends on particle size, which is unknown. A first order approximation of this is to include the wavelength vector λ, with polynomial terms, e.g.:

$$\mathbf{m} = a_i + b_i\mathbf{z}_i + c_i\mathbf{z}_i^2 + d_i\lambda + e_i\lambda^2 + \boldsymbol{\epsilon}_i \tag{9}$$

The parameters a_i, b_i, c_i, d_i, and e_i may be estimated by some sort of linear regression that makes the residual elements in $\boldsymbol{\epsilon}_i$ small, with one separate model for each sample $i = 1, 2, \ldots$. To ensure statistical and numerical stability, regression on standardized regressors was used, with a small ridge parameter.[1]

The primary purpose of the EISC extension terms $\mathbf{z}_i^2$, λ, and λ^2 in Eq. 9 is to improve the estimation of the basic interference effects, the additive offset a_i (reflecting "baseline differences") and the multiplicative slope b_i (reflecting "relative scatter coefficient differences"). But the extensions may also be used explicitly in the subsequent correction. Depending on whether or not one expects the corresponding coefficient estimates c_i, d_i, and e_i to carry information about the analyte, one may choose whether or not to use them in the subsequent correction. If they are thought (or found) to pick up irrelevant complexity from the data, the subsequent calibration modeling may be simplified after the EISC correction:

$$\mathbf{z}_{i,\text{corrected}} = a_i + b_i\mathbf{z}_i + c_i\mathbf{z}_i^2 + d_i\lambda + e_i\lambda^2 \tag{10}$$

This is the EISC correction used in the present paper. Alternatively, if the extension coefficients d_i and e_i for the wavelength are expected to have picked up variation in the analyte that one does not want to lose, the effects $d_i\lambda$ and $e_i\lambda^2$ may be retained in the spectra by reducing the correction to

$$\mathbf{z}_{i,\text{corrected}} = a_i + b_i\mathbf{z}_i + c_i\mathbf{z}_i^2 \tag{11}$$

Note that Eq. 8 is still a simple, linear (additive) model, but Eq. 9 works as a mixed additive/multiplicative preprocessing, in the sense that b_i is a multiplier that may reflect the relative scatter coefficient, while a_i may represent its additive baseline offset. In that sense the EISC correction (Eq. 10 or 11) is analogous to the correction by MSC (Eq. 2) and its extension.[1]

MATERIAL AND METHODS

Samples. Wheat kernels (415) representing 43 different varieties or variety mixtures from two different locations in Denmark made up the calibration set, while wheat kernels (108) representing 11 different varieties from one location made up the test set.[11] All kernels were randomly chosen from bulk samples. The test samples were acquired with the calibration samples, but stored for about two additional months before measurement in order to provide a check for temporal drift in the samples and instrumentation. The NIT single seed data set is made available on the World Wide Web (Pedersen, Pram Nielsen, Munck & Engelsen, NITSingleSeed, www.models.kvl.dk).

Spectra Recordings. The single kernel transmittance spectra were collected on an Infratec 1255 Food and Feed Analyzer (Tecator AB, Höganäs, Sweden). Each kernel was placed in a single seed sample cassette, and transmittance spectra in the range 850–1050 nm were recorded. A tungsten lamp (50 W) and a diffraction grating were used to create monochromatic light. The light passed through the kernel, reaching the silicon detector in a diffuse pattern. Spectra were recorded three times for each kernel and the average of the three spectra was used for the calibrations. The time required for scanning (single scan) 23 single kernels in the cassette was about 90 s.

Protein Determination in Single Kernels. After the spectral recording of the intact wheat kernels each kernel was crushed in the single kernel characterization system (SKCS 4100, Perten Instruments Inc., Reno, NV) and the moisture content necessary for calculation of protein content in dry matter was determined. Subsequently, single kernel nitrogen content was determined directly by a modified Kjeldahl method.[12] Nitrogen in single kernel grits was transformed into ammonium sulfate by digestion (410 °C for 1 h) with 6 mL sulfuric acid (98%). The solution was then alkalized (25 mL 35% NaOH and 75 mL H_2O) and distilled into 25 mL boric acid (0.2%) with methyl red and bromcresol green indicator. The amount of resulting ammonia produced was determined by titration (0.0050 M HCl). The method is based on the assumptions that proteins contain 16 percent nitrogen and that non-protein nitrogen content can be neglected. The protein content is reported as 5.7 times the total nitrogen content for wheat kernels. This unusual calculation factor is due to the high nitrogen content of glutamine. Based on previous experience with samples of 30–40 mg of wheat flour, the analytical error of the analyte was expected to have an absolute standard uncertainty of 0.16% (percent protein content in dry matter).

Data Analysis. Multivariate data analysis was carried out using The Unscrambler version 7.6 (www.camo.com), except for the EISC calculations, which were programmed and carried out using MatLab version 6.1 (The Matworks, Inc., Natick, MA). Conventional multivariate calibration models were developed from the 415 calibration samples using PLSR for protein content (**y**) from NIT spectra (**X**) after different types of spectral pre-transformation (MSC, basic EISC (Eq. 6), EISC with physically extensions (Eq. 10), second derivatives, and combinations of MSC and second derivatives). Optimal numbers of PLSR components (PCs), A_{Opt}, as well as apparent root mean square error of Y-prediction, RMSECV, were estimated by cross validation within the calibration set. To ensure robust and representative segmentation in the cross validation, the 415 calibration samples were sorted for increasing value of protein content (**y**), and then split systematically into 10 cross-validation segments. Performance of calibration models was validated by predicting the protein content in the 108 samples (validation set), yielding the root mean square error of Y-prediction, RMSEP.

RESULTS AND DISCUSSION

Protein Content. The statistics of the Kjeldahl protein determination of the two sample sets are listed in Table I. The protein concentration in the calibration set ranges from 6.8 to 15.2%, while the concentration in the test set ranges from 7.0 to 17.0%. As indicated by the standard deviations in Table I, relatively few kernels have extreme protein content; however, a certain degree of extrapolation is required for the PLSR calibration model to cover the protein range of the test set. The higher protein content in the test samples is probably the result of a certain loss of moisture during the additional storage period.

TABLE I. Means and standard deviations (SD) of single kernel protein data in the calibration and test sets.

Sample set	# of kernels	Mean protein (%)	SD (%)	Min. (%)	Max. (%)
Calibration	415	10.0	1.56	6.8	15.2
Test	108	9.8	1.75	7.0	17.0

Near-Infrared Transmittance Spectra. The NIT spectra of the single wheat kernels presented in this study cover the spectral region from 850 to 1050 nm in 2 nm steps containing primarily the second overtones of O–H (carbohydrates and water) and N–H (protein) stretching vibrations and the third overtone of the C–H (fats) stretching vibration. The fundamental O–H stretch for hydrogen bonded systems is typically found between 3400 and 3300 cm^{-1} (IR), corresponding to 2940–3030 nm, which will ideally give second overtones in the NIR region 980–1010 nm. Secondary amides (proteins) give rise to a fundamental N–H stretching vibration located near 3300 cm^{-1} (IR), corresponding to an ideal second NIR overtone near 1010 nm. Aliphatic C–H stretching vibrations are located between 3000 and 2840 cm^{-1}, corresponding to 3333–3521 nm, which will ideally give rise to third overtones in the Vis/NIR region between 833 and 880 nm. This spectral region is thus of utmost importance to food-related samples, as most of the important functional components are represented. The electromagnetic radiation is relatively high in energy, yet still absolutely nondestructive. Moreover, the absorption of the second and third overtones is much lower than the fundamental and first overtone vibrations, enabling larger sample volumes to be measured, which is very important when measuring heterogeneous systems.

Figure 2 displays raw NIT absorbance spectra of the 415 samples in the calibration set. From the figure, large additive offset and multiplicative scaling effects are readily observed. These are probably due to dispersive loss of light and changes in optical path length, caused by variations in kernel size and texture as well as kernel orientation in the sample cassette. These observed differences in light lost due to physical effects probably overshadow the absorbance changes due to concentration variations in the chemical constituent like starch, water, and protein.

Figure 3A displays the same NIT spectra after EISC pre-transformation according to Eq. 10. In comparison, Fig. 3B shows MSC pre-transformed spectra (Eq. 2), Fig. 3C shows second derivatives of the spectra followed by MSC, and Fig. 3D shows MSC-transformed spectra followed by second derivatives. Compared to the raw spectra in Fig. 2, all the calibration samples appear almost identical after the EISC. However, Fig. 4A shows that after mean centering to remove the average spectral pat-

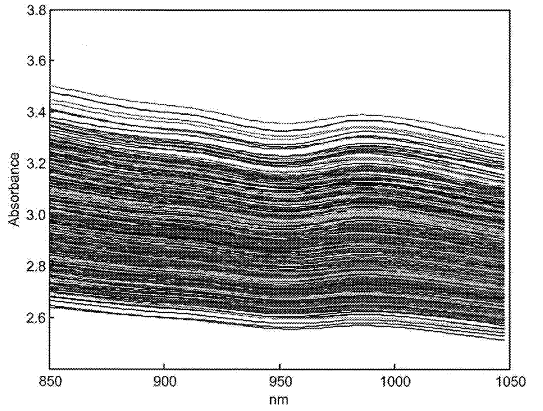


FIG. 2. Raw NIT spectra (850–1050 nm) of the 415 wheat kernels from the calibration set.

tern in the NIT data, the EISC pre-transformed spectra are quite different. Likewise, the mean-centered MSC pre-transformed spectra (Fig. 4B), mean-centered second derivatives of the spectra followed by MSC (Fig. 4C), and mean-centered MSC-transformed spectra followed by second derivatives (Fig. 4D) show clear differences between the samples. The question is whether these differences relate to variations in the protein content.

Figure 5 compares the calibration models from the raw (dotted) spectra (Fig. 2) and the EISC-transformed (solid) spectra (Fig. 3A), both after mean centering. It shows the regression coefficient summary for the two models obtained at a conservative model rank (i.e., 9 and 6 PCs, Fig. 5A) and at the number of PCs that appeared to be near optimal (11 and 7 PCs, Fig. 5B) judging from the cross validation. In general, the EISC has reduced the number of PCs required. Particularly in Fig. 5B, the two predictors are relatively similar, although some differences can be observed. At the slightly lower rank in Fig. 5A the models are even more distinct.

Figure 6 compares the apparent performance of the raw NIT data and the EISC pre-transformation to various other pre-transformations for the single seed protein calibration models. The prediction errors (RMSECV) are plotted against the number of PLSR components. The figure reveals a significant reduction in the number of PLSR components needed, from the raw spectra to the pre-transformed spectra. Secondly and most interestingly, the plot reveals that only the EISC pre-transformation (solid) and the second derivative followed by MSC pre-transformation (densely dotted) are able to provide an optimal model according to the level of the prediction error in the calibration set.

The prediction error in the calibration set for the PLSR model based on EISC-transformed spectra was estimated by cross validation to 0.49% protein (7 PCs). The corresponding prediction error for the PLSR model based on the second derivatives followed by MSC-transformed spectra was also estimated to 0.47% protein (5 PCs), while PLSR models based on raw, differentiated, basic EISC or MSC-treated spectra never reached a prediction error less than 0.55% protein, regardless of the number of PLSR components applied. The improved prediction performance agrees with the findings of Delwiche,[10] who showed that the two-step procedure of using second derivatives followed by MSC gave a better single seed NIT model for prediction of protein content. The most significant result of this comparison is that the single-step EISC performs equally as well as the double transformation, but it is perhaps also noteworthy that basic EISC (ISC) performs just as the sister algorithm MSC on the calibration set, but with a significantly better result on the test set.

The two-step method based on second derivatives fol-

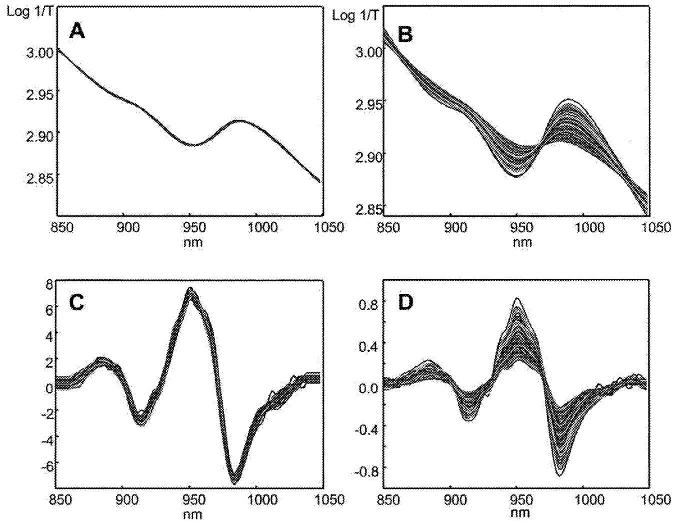


FIG. 3. NIT spectra (850–1050 nm) of the 415 wheat kernels from the calibration set; (**A**) EISC transformed (Eq. 10), (**B**) MSC transformed (Eq. 2), (**C**) second derivatives followed by MSC, and (**D**) MSC followed by second derivatives.

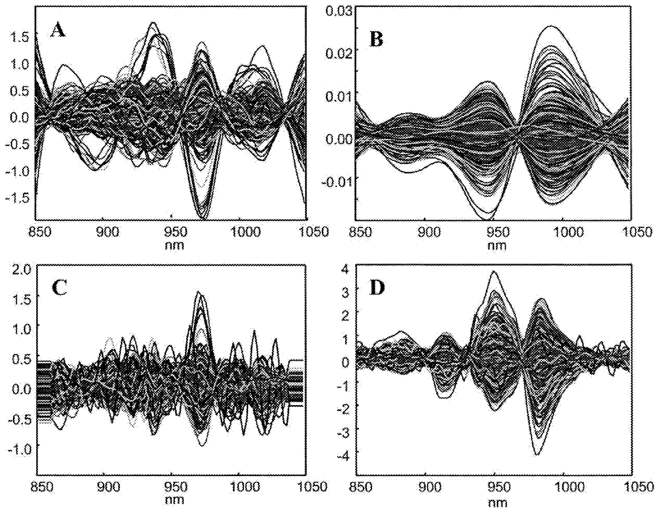


FIG. 4. Mean-centered NIT spectra (850–1050 nm) of the kernels from the calibration set; (**A**) EISC transformed, (**B**) MSC transformed, (**C**) second derivatives followed by MSC, and (**D**) MSC followed by second derivatives.

lowed by MSC-corrected spectra performs considerably better than the opposite two-step method, MSC followed by second derivatives (Fig. 6). This emphasises that the order of the applied pre-transformations is important and that conventional MSC is a poor model when the scatter is not linear.[13] In the MSC it is assumed that the scatter is linear throughout the spectral range, since the whole spectrum is linearly adjusted by one slope and one offset. However, if the loss of light due to light scattering and other effects is not this simple, the MSC correction is not suitable and, consequently, a model based on MSC followed by the second derivatives will not be optimal. On the contrary, by using the second derivatives, it appears that the spectra are successfully corrected for local offsets and linear trend variations.

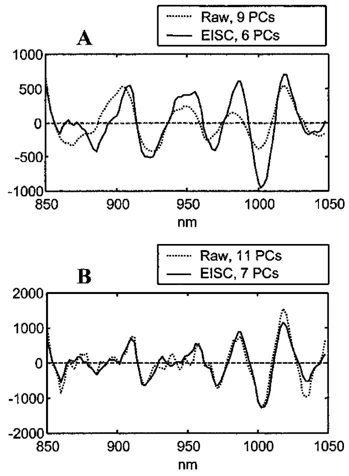


FIG. 5. The regression coefficients for the calibration of raw input spectra (········) for 9 PCs (**A**) and the optimal 11 PCs (**B**), and for the calibration of EISC-transformed spectra (———) for 6 PCs (**A**) and the optimal 7 PCs (**B**).

Validation. The test set (108 single wheat kernels representing 11 of the varieties included in the calibration, but stored for an additional two months) was measured on the same NIT instrument, analyzed for protein content by the same method, and used for testing the (long-term) stability of various calibration models. Figure 7 shows the EISC parameter estimates according to Eq. 9, the coefficients a_i (additive offset), b_i (multiplicative scaling), c_i (for squared spectrum), d_i (for wavelength), and e_i (for squared wavelength) for the 415 calibration samples as well as for the 108 test samples. Figure 7F illustrates that the samples have been sorted for increasing protein content (**y**) within each of the two sample sets for simpler

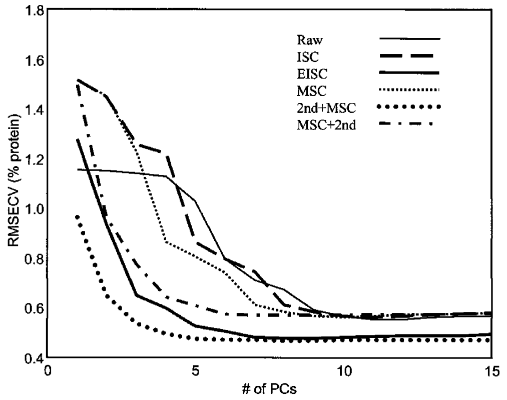


FIG. 6. RMSECV vs. PLSR components (PCs) for models based on different pre-transformed NIT spectra; raw spectra, ISC-corrected spectra, EISC-corrected spectra, MSC-corrected spectra, spectra transformed to the second derivative followed by MSC correction (2nd + MSC) and MSC correction followed by second derivative (MSC + 2nd).

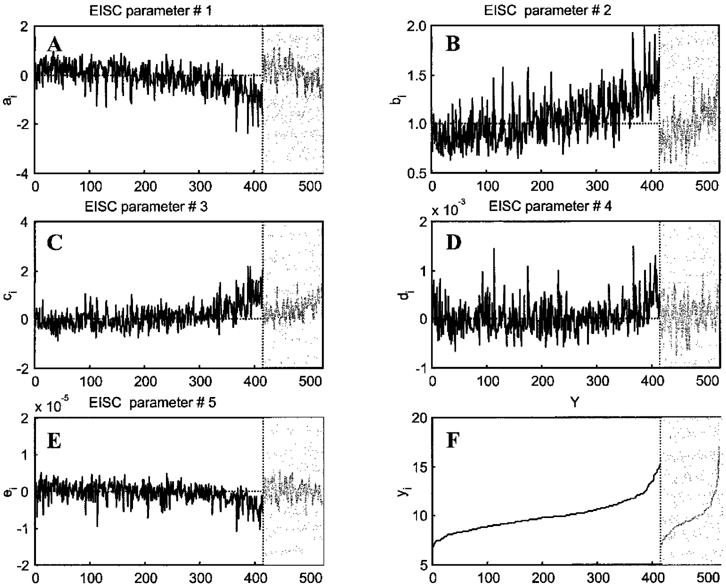


FIG. 7. The estimated EISC parameters (Eq. 9) plotted against sample *i* for the 415 calibration samples (black) as well as for the 108 test samples (grey), sorted according to increasing protein content in the two data sets; (**A**) EISC parameter # 1: a_i (additive/offset); (**B**) EISC parameter # 2: b_i (multiplicative/relative scatter scaling of input spectrum); (**C**) EISC parameter # 3: c_i (effect of squared spectrum); (**D**) EISC parameter # 4: d_i (effect of wavelength); and (**E**) e_i (effect of squared wavelength). The increasing protein content y_i within each of the two sample sets is shown in subplot **F.**

cross validation in the calibration set and simpler visual interpretation of the EISC parameters in both sets. The EISC parameters (Figs. 7A–7E) show highly erratic variations, especially in b_i (multiplicative scaling) and d_i (wavelength). But some systematic changes with the protein content may be observed in both sets, particularly at the highest protein levels (>11%). This is an indication that the EISC may have picked up and removed some variation related to the analyte. The cause and nature of this lost analyte information is unclear, but needs to be studied in more detail.

Figure 8 compares the calibration set and the test set, before and after the EISC. Spectroscopically, the mean-centered spectra of the test samples appear normal in the raw data (Fig. 8B), compared to the calibration set (Fig. 8A), while they show a very distinct pattern after the EISC (Fig. 8E vs. Fig. 8D). This is an indication that all test samples deviate in the same systematic way from the calibration mean spectrum **m.** If this type of deviation is also present among the calibration samples, it may be modeled and corrected for in the calibration model; if not, the systematic deviations will cause grave errors in the predicted percent protein in the test set. The peaks just below 950 nm in Fig. 8D indicate that some of the calibration samples indeed display the same general pattern, but this needs to be verified in the prediction of protein.

Figures 8C and 8F compare the predictive performance before and after EISC. The long curves show the estimated error for protein content **y** predicted from the 100 NIT wavelength channels **X** for PLSR calibration models using between 0 and 15 PCs, for the cross-validated calibration set (RMSECV, solid) and the test set (RMSEP, dashed). When using the raw spectra in Fig. 8C as **X,** the cross validation shows that several of the first PCs (2, 3, 4, 5) have little or no predictive relevance for the protein content; hence, they must reflect very strong covariance structures in the NIT spectra. More importantly, the predictive ability in the test set changes erratically with the increasing number of PCs; obviously, a wrong choice in the number of PCs to be used for prediction may cause very high prediction errors in the test set.

In contrast, when using the spectra after EISC in Fig. 8F as **X,** the cross-validation curve falls smoothly, as desired. The model is shown to require at least 4 PCs. The test set curve is very similar to the cross-validation curve after 4 PCs.

The two short curves in Fig. 8F show the estimated prediction errors using instead the 5 EISC parameters $[a_i, b_i, c_i, d_i, e_i]$ (Eq. 9) from the different samples as **X,** instead of the 100 wavelength channels. Some predictive ability for the protein content (**y**) is evident in the cross-validation curve (squares). Hence, the EISC may have removed Y-relevant information. However, in the test set (diamonds) the predictive ability for **y** is not as good. Attempts (not shown here) at joining the NIT data with the EISC parameters as extra variables, $\mathbf{X} = [\mathbf{z}_{i,\text{corrected}}, a_i, b_i, c_i, d_i, e_i]$, using weighted least-squares PLSR, gave a slight but insignificant improvement in RMSECV (calibration set), but no improvement in RMSEP (test set). It thus appears that the "physical" information apparently removed by the EISC in these data was not important or reliable for the prediction of chemical protein content: the errors that they contribute to the calibration model

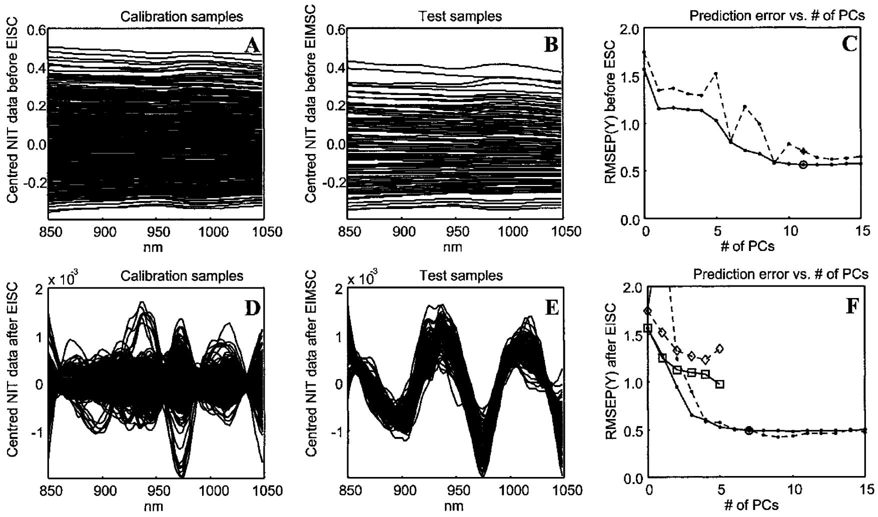


FIG. 8. The mean-centered NIT spectra for (**A**) the calibration samples and (**B**) the test samples, and the EISC transformed mean-centered NIT spectra for (**D**) the calibration samples and (**E**) the test samples. Prediction error vs. the number of PCs for the calibration samples (———) and the test samples (– – –) (**C**) before EISC transformation and (**F**) after EISC transformation. The two short curves in subplot **F** show the prediction errors calibrating only with the five EISC parameters (Eq. 9), $\mathbf{X} = [a_i, b_i, c_i, d_i, e_i]$: the calibration samples (———) and the test samples (– – –).

are greater than the otherwise unmodeled Y-variation that they can remove.

The calibration and the test set results for all the tested pre-transformation methods are summarized in Table II in terms of the RMSECV (calibration set) and RMSEP (test set) read at the optimal number of PCs, and of the correlation coefficients based thereon. Compared to the untransformed raw data, the basic EISC/ISC did not affect the results very much. However, there is an improved correlation, both for the calibration set (from 0.93 to 0.95) and for the test set (0.96 to 0.98), after applying the full EISC (Eq. 10) to the NIT spectra. The protein calibration model predicts the test kernels well throughout the protein range; the prediction error (RMSECV and RMSEP) ends as low as 0.49% protein in a protein range of 7 to 17%. This RMSE level approaches the sampling and measurement error on the single seed protein determination (0.16% determined for samples of 30–40 mg of flour), and the results demonstrate a very good and robust protein calibration on single wheat kernels.

TABLE II. Performance statistics of the PLSR models for single seed protein predictions using single seed NIT spectra from the calibration set (415 kernels) and the subsequent test set (108 kernels). CV is cross validation. RMSECV is the root mean square error of cross validation, and RMSEP is the root mean square error of prediction.

		Correlation		Prediction error (% protein)	
Pre-transformation	# of PLSR components	Cal. set (CV)	Test set	Cal. set RMSECV	Test set RMSEP
Raw	11	0.93	0.96	0.55	0.70
ISC	9	0.93	0.95	0.58	0.69
EISC	7	0.95	0.98	0.49	0.49
MSC	9	0.93	0.95	0.57	0.78
2nd + MSC	5	0.95	0.98	0.47	0.48
MSC + 2nd	7	0.92	0.93	0.60	0.66

The conclusion is that the basic ISC performs equally as well as the traditional MSC, perhaps even slightly less aggressively on the calibration set, resulting in an improved test set prediction. Both the EISC with general (physical) extensions and the two-step "second derivatives followed by MSC" in this data set can correct for spectra interferences that are not corrected by the more "classical" pre-transformations, MSC or second derivatives. The EISC is particularly promising because it is more flexible and easier to understand than the "classical" MSC and two-step methods. In this study we have emphasized a general applicable version of the EISC, but its flexible approach allows simple implementation of system-specific interferences such as known analytes.[1] In a future implementation, we will work on a version in which the correction coefficients are constrained to be orthogonal to the reference value **y**, with the aim of optimizing subsequent regression models with even less loss of analyte information.

ACKNOWLEDGMENTS

The authors wish to thank Danish Bacon and Meat Council and the Ministry of Food, Agriculture and Fisheries for financial support through the project "Early post mortem measurement of WHC (water-holding capacity) and drip loss in fresh pork" and to the Ministry of Food, Agriculture and Fisheries through the Danish Cereal Network (B4). Gilda Kischinovsky is acknowledged for assistance with the manuscript. The relationshop of the EISC to the following patents is acknowledged: E. Stark and H. Martens: "Multiplacative Signal Correction Method and Apparatus", U.S. Patent 5,568,400; and E. Stark and H. Martens: "Improved Multiplicative Signal Correction Method and Apparatus", European Patent 0415401.

1. H. Martens, J. P. Nielsen, and S. B. Engelsen, Anal. Chem., paper submitted (2002).
2. H. Martens and T. Næs, *Multivariate Calibration* (John Wiley and Sons, New York, 1993).
3. S. Wold, H. Martens, and H. Wold, *Lecture Notes in Mathematics,*

A. Ruhe and B. Kågström, Eds. (Springer Verlag, Heidelberg, 1982), p. 286.

4. H. Martens and M. Martens, *Multivariate Analysis of Quality. An Introduction* (John Wiley and Sons, Chichester, 2001).
5. I. S. Helland, T. Næs, and T. Isaksson, Chemom. Intell. Lab. Syst. **29,** 233 (1995).
6. P. Geladi, D. McDougall, and H. Martens, Appl. Spectrosc. **39,** 491 (1985).
7. H. Martens, S. A. Jensen, and P. Geladi, "Multivariate Linearity Transformation for Near-Infrared Reflectance Spectrometry", in *Symposium on Applied Statistics,* O. H. J. Christie, Ed. (Stokkand Forlag Publ. Nordic., Stavanger, Norway, N-4000, 1983), p. 208.
8. H. Martens and E. Stark, J. Pharm. Biomed. Anal. **9,** 625 (1991).
9. J. Kjeldahl, Anal. Chem. **22,** 366 (1883).
10. S. R. Delwiche, Cereal Chemistry **72,** 11 (1995).
11. J. P. Nielsen, D. K. Pedersen, and L. Munck, Cereal Chemistry, paper submitted (2002).
12. American Association of Cereal Chemists, AACC Method 46-12. Crude Protein—Kjeldahl Method, Boric Acid Modification. *Approved Methods of the American Association of Cereal Chemists* (1995).
13. H. Martens and T. Næs, "Multivariate Calibration by Data Compression", in *Near-Infrared Technology in the Agricultural and Food Industries,* P. C. Williams and K. H. Norris, Eds. (American Association of Cereal Chemists, St. Paul, 2001), 2nd ed.