2009

Intra and inter-brand calibration transfer for near infrared spectrometers

Benoit Igne
Iowa State University

Intra and inter-brand calibration transfer for near infrared spectrometers

by

Benoit Igne

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial and Agricultural Technology

Program of Study Committee:
Charles R. Hurburgh Jr., Major Professor
Thomas J. Brumm
Brian L. Steward
Philip M. Dixon
Stephanie Jung

Iowa State University

Ames, Iowa

2009

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The completion of this dissertation would never have been possible without the interactions I have had during these past three and a half years with my major professor, Dr. Charles R. Hurburgh Jr., with staff and graduates students of the Grain Quality Laboratory, especially my office mates. I would like to thank member of my program of study committee, Dr. Brian L. Steward, Dr. Thomas J. Brumm, Dr. Philip M. Dixon, Dr. Stephanie Jung, and Dr. Carl J. Bern.

I also would like to thank people external to Iowa State University with which I have collaborated on this dissertation project and various research programs, Dr. Sylvie Roussel, Dr. Fabien Chauchard, Dr. Jean-Michel Roger, Dr. Veronique Bellon-Maurel, Dr. Roumiana Tsenkova, Dr. James Reeves, Dr. David Funk, Dr. Frederic Violleau, Cecile Levasseur, and Dr. Igor V. Kovalenko.

I extend many thanks to Dr. Tom Fearn and Dr. Joan Trygg for their help in implementing some of the algorithm used in this study.

Finally, I would like to thank Courtney White for her help and support during those long hours spent away from her, working on completing my gradate education. I would also like to thank my parents and family for their support during these three and a half years spent away from home.

This project was funded by the United Soybean Board and carried out in the Grain Quality Laboratory at Iowa State University.

ABSTRACT

Robust modeling methods were implemented for the transfer of near-infrared calibration models in intra and inter-brand situations. A network of four instruments from two brands (Foss Infratecs and Bruins OmegAnalyzerGs) was used to implement spectral pretreatment methods, local and variable selection techniques, and orthogonal methods to transfer protein, oil, and linolenic acid models across instruments of the same brand and across instruments of different brands. A total of fifty seven techniques were implemented among which spectral filtering methods based on the smoothing of high frequency components of Fourier and wavelet transforms. A new approach to local similarity was introduced. Results showed that the effectiveness of the various methods was instrument, parameter, and validation set dependent. In some situations, no differences could be observed between master and secondary unit predictions. Local methods appeared to be the weakest methods, most likely due to a problem of over-fitting (specialization) of the calibration set. The transfer of calibrations across brands was possible with performances similar, or better, than in intra-brand calibration transfer.

CHAPTER 1.

GENERAL INTRODUCTION

Today's challenges for food quality and traceability have created the need for fast, inexpensive, and easy to use methods to control and certify industrial, pharmaceutical, and agricultural products. In the grain industry, near-infrared spectroscopy has become a tool of choice for screening, controlling, and assessing quality. It is an inexpensive analytical method based on the absorption of near-infrared light by organic molecules proportionally to their distribution. Near-infrared spectroscopy has been successfully used in many situations (i.e. proximate analysis, fatty acid, and amino acids profile determinations).

The development of a near-infrared application requires the constitution of a calibration (or prediction) model which will be used to interpret absorption spectra from samples to analyze. Through a calibration process, the relationship between absorption spectra and the compound of interest is established under the form of a linear or non-linear regression called a prediction model. After validation, these models are used in routine operations. In many cases, it is necessary to calibrate a network of instrument within the same factory or location and it is often not economically viable to recalibrate each instrument. The implementation of methods allowing the transfer of calibration models from a master unit to operational units is then necessary.

Three approaches exit for calibration transfer: optical methods that aim to modify spectra of the secondary units to match those on the master unit, post-regression correction methods that correct predicted values of the master's model on the secondary unit to match master's predictions, and robust methods that tune the calibration set to perform well in encountered situations. Robust modeling is a very interesting and satisfactory approach to calibration transfer since it aims to remove noise responsible for poor performances from the information available. In calibration transfer situations, inter-instrumental differences is what will most likely to be removed.

The literature provides numerous applications of optical and post-regression correction methods to calibration transfer. Robust modeling has been extensively studied in the last decade, but research available is not focused on calibration transfer. In addition, most of the studies deal with intra-brand calibration transfer while the possibility to transfer historical databases from brand to brand is one of the main issues of today's near-infrared users.

The following dissertation will present, propose, and apply numerous approaches to robust modeling for the development and transfer in intra and inter-brand situations of whole soybean grain for protein, oil, and linolenic acid models. In the second chapter, near-infrared spectroscopy will be introduced in detail and a review of the current calibration transfer methods will be established. The third chapter will present the various standardization methods used in this study, while the fourth chapter will provide results. The last chapter will summarize the results and provide guidelines for robust modeling in near-infrared spectroscopy.

CHAPTER 2.

UTILIZATION OF NEAR INFRARED SPECTROSCOPY FOR THE DETERMINATION OF PRODUCT PROPERTIES OF RAW AND PROCESSED MATTER

The need for traceability is increasingly important in globalized societies where a product can be ordered by a customer in Ames, Iowa and cross the world in plane, ship, or train from Europe, Asia or Africa to be delivered through a postal system employing trucks, cars, and sometime bikes or pedestrians. The traceability of food products is especially important because it directly impacts customers and their health. One certainly remembers the outbreak of Escherichia Coli on spinach in September 2006 where the consumption of bag-conditioned spinach was the source of a massive recall by a California company and was responsible for 199 cases of food borne infections (Food and Drug Administration, 2007). In March 2007, pet-food was contaminated with melamine, a fire retardant, causing the death of 104 animals, mainly from kidney failures (Food and Drug Administration, 2007).

Traceability has thus become an essential element of the control of the food quality. Since 9/11 and the terrorist attacks on North American territory, official acts, such as the PATRIOT Act, clearly state the necessity for companies to record and be able to provide, in a limited period of time, all necessary information to officials (The USA PATRIOT Act, 2005).

The control of the food quality is an essential parameter of traceability and is a critical part of any food product process.

2.1. Control of the grain quality

The world crop production represented 2,115 million metric tons of grain for the crop year 2007 (cotton not included). Corn (37%) and wheat (28%) were the two major field crops, followed by rice (20%), and soybean (10%). The other 5% were cereals such as barley, oats, rye, and other field crops: sorghum, cottonseed, peanut, sunflower seed, rapeseed, and copra. The United States of America (USA) produced 19.6% of the world's crops. Among the major products, the USA produced 42% of the corn, 32% of the soybean, and 9% of the wheat. In addition to the production, the USA exports a significant part of its production (especially soybean, corn and wheat). In 2007 crop year, it exported 61% of its wheat, 44%of its soybean, and 18% of its corn (United State Department of Agriculture, 2007).

The production system as well as the importation/exportation system relies on the measurement of the grade and the quality of the grain at the different components of the chain. As mentioned by Hurburgh and Brumm (2004), "Description (grading) of grain by quality characteristics makes possible the trading of grain contracts with neither buyer nor seller personally inspecting the grain being traded". The USDA has established a grading system based on the density, the quantity of broken grains, the presence of foreign material, and the grain moisture. In addition to the grade, a lot may be tested for additional parameters (protein, oil, starch, fiber …) to meet the quality standard of a specific application. In the grain system, as summarized in the figure 1, various analyses are performed along the chain.

At the field level, the moisture is the principal parameter of quality that is controlled. Harvesting a grain too humid will require additional drying expenses either at the farm storage unit or at the elevator to maintain its quality during storage. Thus, farmers use moisture meters, usually based on conductance, that measure the ease with which electric current flows through the grain. At the elevator, the grade of the grain is determined and lots of similar qualities are accumulated into large silos. The determination of the grade involves the use of a Carter Dockage Tester (US Patent #1644872) for the determination of

damaged grains and foreign materials. Test weight and moisture are often determined by a single instrument based on conductance. Analytical instruments are approved by the national type evaluation program (NTEP) for use in the determination of the grade. From the elevator to the food companies or to exportation, additional information on the quality of the grain may be asked by buyers. In these cases, non-grade tests such as near infrared spectroscopy (NIRS) may be used to determine, for instance, protein, oil, fiber, and fatty acid profile for soybeans.

NIRS is increasingly involved in the process of determination of the grain quality in the grain chain because several manufacturers are investigating the use of NIRS, in addition to spatial data, to model in real time the performances of a field (Brumm, 2007). In addition to the determination of the grade and the composition of the grain, the need for rapid and inexpensive methods for food safety screening of grain and foodstuff lots is becoming important. NIRS has shown interesting results in the determination of the contamination of grain by mold using ergosterol, a biomarker present in the nuclear membrane and vacuole (Levasseur, 2004) or in the quantification of the mycotoxical contamination (Kim et al, 2004; Berardo et al, 2005). The melamine in pet food was detected by NIRS, identifying that the nitrogen content was not from a protein form (Marcott, 2008).

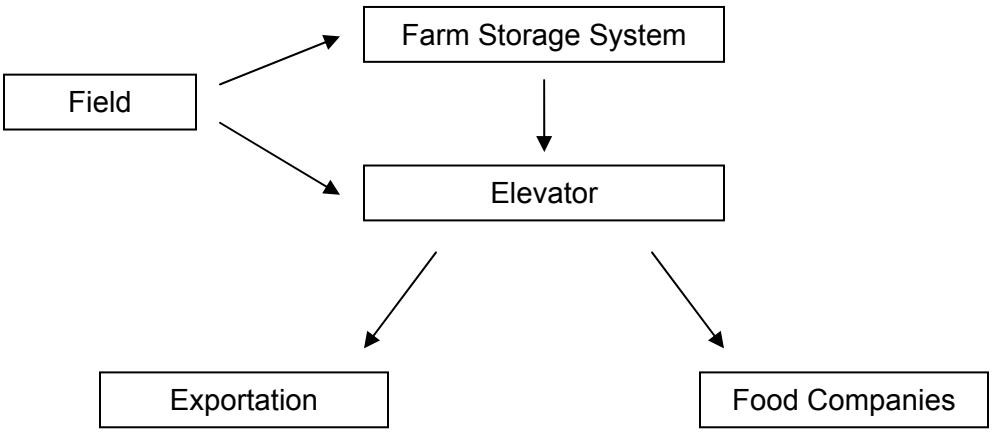


Figure 1: Summary of the grain system, from the field to the fork.

2.2. Near infrared spectroscopy

Near infrared spectroscopy is widely used because of its ability to provide an inexpensive, fast, and real time control of industrial processes. It has shown its usefulness in the analysis of agricultural products and foodstuffs, polymers, textiles, pharmaceutical and medical sciences, soil, and petrochemicals (Williams and Norris, 2001; Siesler et al, 2002; Burns and Ciurczak, 2001;).

2.2.1. Theoretical introduction

Near infrared spectroscopy is a type of spectrochemical analysis that uses the near infrared part of the electromagnetic spectrum (770 – 2,500 nm or 1.2×10^{14} – 3.9×10^{14} Hz) to obtain information on the analyzed matter based on its absorption of the incident light (Ingle and Crouch, 1988). For absorption to occur, the frequency of the incident radiation must correspond to the energy difference between two energetic states of a chemical bound. Absorption is defined as

$$A = -\log T = -\log \frac{\Phi}{\Phi_0} \tag{1}$$

where **A**, the absorbance, is a function of **T,** the transmittance, that corresponds to the ratio of the radiant power transmitted, **Φ,** to the incident radiation, $\mathbf{\Phi_0}$. Figure 2 illustrates absorption in the basic setup of a transmission near infrared spectrometer.

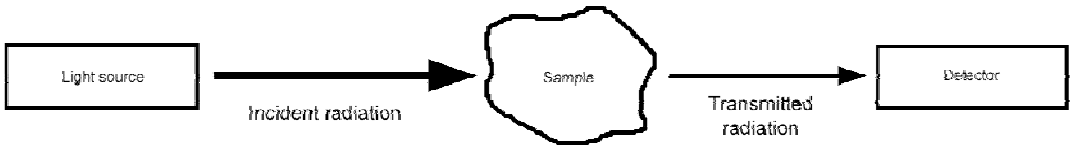


Figure 2: Simplified setup of a near infrared transmission spectrometer. The ratio transmitted radiation to incident radiation defines the transmittance of the sample.

The absorption of infrared energy causes bonds to vibrate and rotate. The vibration of a dipole can be seen on figure 3.A. It represents a diatomic oscillator that vibrates based on the energy level at which the dipole is. This theoretical model is called harmonic oscillator and can be expressed as

$$V = \frac{1}{2}kx^2 \tag{2}$$

where **V**, the potential energy of the vibrating system, is a function of **k** and **x**, atom's equilibrium position and the displacement of atoms from their equilibrium position respectively.

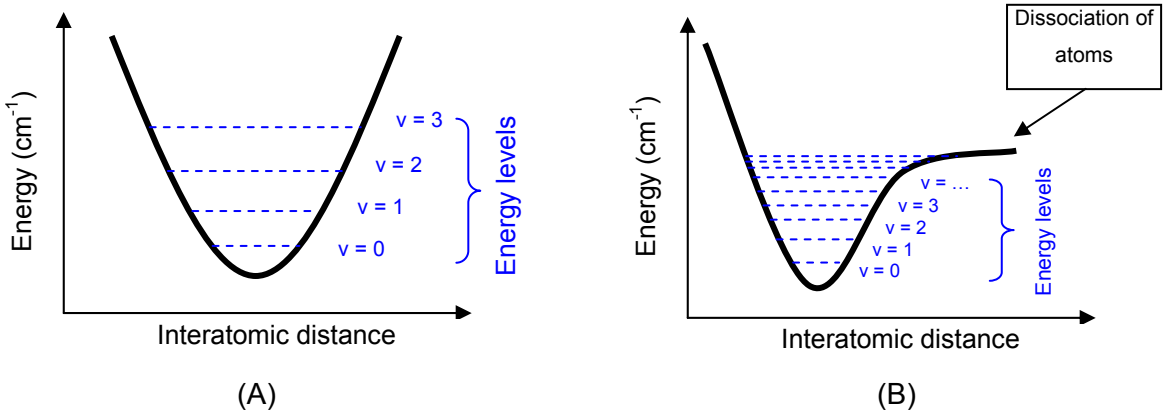


Figure 3: Vibrational energy levels of the harmonic model (A) and the anharmonic model (B). Modified from Kovalenko, 2006.

When an infrared beam is sent to the dipole, the energy level is raised and the dipole oscillates in the limits of the energy levels. When a dipole receives enough energy to raise its energy level from the ground vibrational level to the first vibrational energy level (defined by quantum energy), the molecule experiences a fundamental transition. Fundamental transitions are observed in the mid infrared region (2,500 – 50,000 nm) (Bokobza, 2002).

The harmonic oscillator is not sufficient to explain all responses encountered in infrared spectroscopy because if only fundamental bands were existing, no absorption would be observed in the near infrared region. The ideal harmonic model is modified by two behaviors of real molecular vibrations: mechanical anharmonicity and electrical

anharmonicity. Mechanical anharmonicity is a consequence of the structure of atoms where nuclei, when close to each other, experience a strong repelling force that can lead to dissociation of the bond. In this situation, the displacement of atoms can be approximated by a higher-order equation than equation 2, such as the anhormonic oscillator proposed in figure 3.B. The electrical anharmonicity is a consequence of the non-linear relationship between dipole moment and atomic displacement. For a heteronuclear diatomic molecule, the maximal dipole moment is observed at an atomic separation slightly smaller than the equilibrium atomic separation (Miller, 2001).

This anharmonicity is at the origin of phenomena encountered in the near infrared region: overtone transitions, combination modes, and unequal distance between vibrational energy states. Overtone absorptions correspond to energetic changes where the molecule experiences an increase of its vibrational energy level to the second, third, or the fourth level. Raises to the fifth and higher vibrational energy levels are possible, but the energy absorption would be too weak to be useful for practical analysis (a consequence of the unequal distance between vibrational energy states). It is important to note that the energy needed to raise the vibrational energy level from the fundamental state to first or second overtones is between 10 and 1000 times weaker than the energy needed for the fundamental transition. The energy needed to raise the vibrational energy level to higher states corresponds to frequencies in the short wavelength of the mid infrared or in the near infrared region. "A fundamental carbonyl […] vibration at 1,750 cm^{-1} or 5,714 nm would have a first overtone at approximately 3,000 nm, a weaker second overtone at 2,100 nm, and a third, very weak overtone at 1,650 nm. The fourth overtone, at about 1,370 nm, would be so weak as to be analytically useless." (Ciurczak, 2001). Absorption bands are often expressed in reciprocal centimeters (cm^{-1}) that are directly proportional to the energy level. The conversion between nanometers and reciprocal centimeters is as follows

$$wavelength_{nm} = \frac{1}{wavenumber_{cm-1}} \tag{3}$$

In addition to the ability of an absorption band to be produced at two or three times the frequency of the fundamental vibration, vibrations have a tendency to combine to give a single absorption band by addition and subtraction of transition frequencies. Table 1 presents a classical example of fundamental, overtones, and combinations frequencies for the sulfur dioxide molecule (Bokobza, 2002).

Table 1: Assignment of some absorption bands of the sulfur dioxide molecule (modified from Bokobza, 2002).

Vibrational frequency (nm)	Assignment
19,268	v2
16,502	v1 - v2
8,688	v1
7,347	v3
5,344	v2 + v3
4,355	2v1
4,002	v1 + v3

Thus, a molecule will present absorption bands from not only the fundamental transition (v1) but also from overtone transitions (v2 and v3) and some combination bands (v1 - v2, v2 + v3, and v1 + v3).

The vibrations of the molecule subject to a raise of its energetic state due to an infrared light can take various forms. Figure 4 shows the different vibration modes that sulfur dioxide can experience, depending on the energy state at which it is. Other vibration modes such as twisting, rocking, and wagging can be observed, depending on the energy state, on the structure of the molecule, and on the phase of the molecule. Rotation phenomena can mostly be observed in gas phase; peak broadening in solid and liquid phases cover the absorption due to rotation.

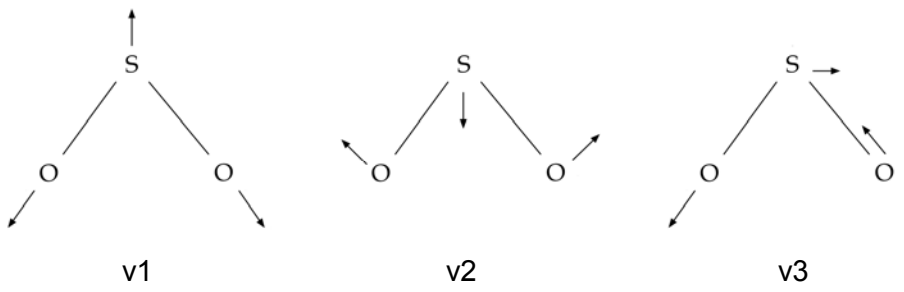


Figure 4: Vibration modes of sulfur dioxide. v1 corresponds to a symmetric stretch at 8,688 nm; v2 corresponds to a bending mode at 19,268 nm; v3 correspond to a antisymmetric stretch at 7,357 nm (Bokobza, 2002).

When the product of interest is mainly formed by C-H, N-H, and O-H functional groups with fundamental absorption between 4,000 and 2,500 nm, the near infrared region is mostly composed of overtones and combinations of these dipoles. Thus, the study of agricultural products is based on the analysis of absorption spectra of overtone and combination bands from C, H, O, and N bonds (Ciurczak, 2001). For a complex mixture, the near infrared spectrum will consist of an overlap of overtone and combination bands from the different functional groups present. A visual interpretation of complex mixture is then made impossible and statistical techniques called chemometrics methods are used to extract useful information. However, depending on the instrumental setup of the spectrometer, the type of analysis can vary.

2.2.2. Near infrared instrumentation

For developers, NIRS consists in collecting the absorbance spectrum of a sample to either compare it to databases in the case of qualitative studies, or perform calculations on absorbance values at each collected wavelength for quantitative applications. The spectrum is the fingerprint of the sample, an indirect indication of its chemical composition. In practice, this technique can appear very simple to end-users. Their role is to present the sample, prepared or not, to the instrument and start the analysis. A short time later (0.1 – 2.0 min), results appear on the screen of the embedded computer and stored. Manufacturers

have been working to improve hardware and software of their instruments to facilitate their use by people without engineering or chemistry backgrounds.

Two types of near infrared spectrometers coexist; units based on transmission and units based on reflection (figure 5.A and Figure 5.B respectively). In transmission, the light is sent through the samples and the detector measures the remaining part of the light (the light intensity at each measured wavelength). A transmittance spectrum is obtained and transformed to absorbance units using equation 1. In reflection, the light is sent to the surface of the sample and a detector measures the light intensity that is reflected. In this situation, only the surface of the sample is scanned and thus requires a high homogeneity. The absorbance spectrum is obtained by changing the transmittance term to the reflectance analog in equation 1. The choice between these two techniques is mainly governed by the type of samples and the application. The treatment of the data remains the same.

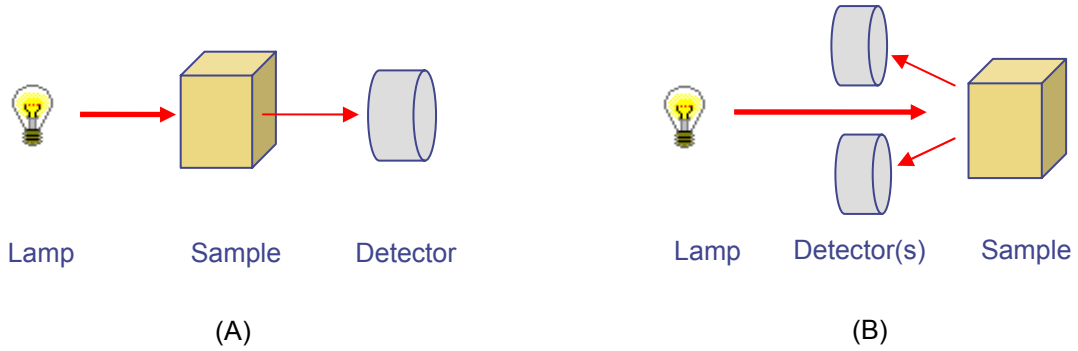


Figure 5: Basic near infrared spectrometer setups. (A) transmittance unit, (B) reflectance unit (modified from Kovalenko, 2006).

In terms of equipment, a spectrometer is globally composed of a light source, a light diffraction system, a sample presentation system, a detector, an analog to digital converter, and an integration board with a firmware or software to appropriately process spectra. Each manufacturer offers different light sources, sample presentation systems, and detection techniques. Ingle et al (1988), Workman et al (2001), Kawata (2002), and McClure (2001) provide comprehensive chapters on these elements. An instrument can be dedicated to analyze absorption of a sample at specific wavelengths (filter instruments) or along a range of wavelengths. The first would output absorption values at selected wavelengths while the second would output absorption values, at each or every two, four, eight wavelengths

(depending on the interpolation). Spectra have high collinearity, which greatly impacts the statistical methods employed.

While the treatment of the spectra is mostly transparent to end-users because performed by a firmware or software, it is one of the most challenging tasks for manufacturers and chemometricians. A spectrometer needs to be calibrated to provide useful information. The process of calibration and its management is, with the design of instruments, the largest area of research in NIRS.

2.2.3. The Calibration Process

The purpose of the calibration process is to establish a relationship between the absorbance spectrum of an "unknown" sample and the element(s) of interest in that sample. In NIRS, the establishment of this relationship requires several steps. First of all, a calibration set needs to be created. This pool of samples must be representative of the variability that the application will encounter. This wide range is necessary to ensure that the calibration model will be able to predict accurately and precisely new samples. The second step is the determination, for each sample selected in the calibration set, of the content in the element of interest by reference analysis. The third step consists of creating the prediction model or calibration. This is done by using multivariate regression methods. Once the prediction model is developed, it is necessary to validate it by using either an independent dataset or a validation technique on calibration samples.

2.2.3.1. Data Collection

Igne et al (2007a) presented a study of the development of a calibration model on a matrix (triticale grain) experiencing large changes from year to year, from breeding and annual variability due to climatic changes. Conclusions were that a constant adaptation of the calibration set must be performed to ensure that the best set of samples is present in the calibration pool to predict new samples. Chemometricians select samples on a yearly or

other periodic basis or by using programs that reevaluate the appropriateness of each sample in the calibration set. The software WinISI (Infrasoft International LLC., State College, PA) for instance, uses principal component analysis and Mahalanobis distance to determine if a new sample is appropriate to be added to the calibration. Samples bringing information already present in the calibration set can be left out.

NIRS is a secondary technique. Performance of the calibration models is based on the quality of the reference measurement. The distribution of the reference values must be carefully controlled. Linear regression techniques (multiple linear regression, principal component analysis, partial least squares regression) can perform on both uniform and normal distributions of the data while non-linear regression methods (artificial neural networks regression and least squares support vector machines) can be successfully applied to datasets presenting other types of distributions. However, in terms of applicability of prediction models to "unknown" samples, prediction models developed on uniform distributions are more likely to provide the best estimation of the true concentration of the compound of interest. Calibration models developed with a normal distribution of the samples are subject to the "Dunne" effect and may cause predictions to regress toward the mean (Williams, 2001).

The data collection procedure should include a detection technique for outliers. Outliers are samples that are different from the population of interest, and that do not represent features of analytical value in the future. This difference can come from hardware failures, a degradation of the sample influencing its chemical composition, or from the fact that the sample belongs to another population and cannot be included in the population of interest. The detection of such samples is typically performed through a principal component analysis with a Hotelling's T^2 test – a multivariate method that tests the membership of an observation to a group – at a given confidence interval. Samples that, during the regression process, present an abnormally large residual value can also be considered for being outliers. It is necessary to remove obvious outliers from the calibration set because they may influence the regression equation and degrade its future performances. However, chemometricians must be careful not to remove too many samples

or relevant samples because those "abnormal" samples can be representative of a new variability that should be included in the calibration set to ensure accuracy and precision in the prediction of future samples (Walmsley, 2006).

2.2.3.2. Signal processing

To maximize the relationship between reference values and spectral data, the signal-to-noise ratio needs to be increased by reducing the noise level. In NIRS, the noise can have several sources: light scattering, spectral interferences (due to overlapping of absorption bands from different functional groups), and spectral distortions (baseline drift, noise in the hardware …).

2.2.3.2.1. Noise and baseline effect

The first type of spectral pretreatment method aims to limit the noise, offset, and baseline effects by removing the high and low-frequency interferences (additional noise and background information) using the collinear nature of the absorbance values. The main techniques are:

(1) Offset correction – It consists of removing a constant from each spectrum

$$X_{ij} = x_{ij} - C \tag{4}$$

where $\mathbf{X_{ij}}$ is the corrected spectrum at i row and j column, $\mathbf{x_{ik}}$ the original spectrum, and **C** the correction factor. This processing implies that the baseline is constant.

(2) Weighted least squares baseline correction – A low order polynomial representing the background is approximated and subtracted to remove the background. It is an iterative method that fits a polynomial to each spectrum in order to approximate the baseline. Highly negative peaks are limited. This method may not apply to situations where all absorption values correspond to

sample absorption (no part of the signal corresponding to baseline only) because it may remove signal and introduce variance in the dataset.

(3) Detrend – It is a non weighted form of baseline removal where a polynomial of a given order is fit to the entire spectrum and subtracted. This technique works only when the background signal is constant, linear, or curved.

(4) Smoothing – It is a low-pass filter (removes high frequency noise). It uses the collinearity between spectral variables to fit an n^{th}-order polynomial to a limited number of spectral points. The larger the window and the lower the order, the bigger the smoothing. Smoothing can be very useful when dealing with noisy signals, but can delete poorly resolved peaks.

(5) Derivative – It removes baseline and enhances peaks. It is often used after a smoothing (to limit the accentuation of noises) by subtracting the x_{ij} absorbance value (i^{th} sample and j^{th} variable) by the x_{ij-1} for a first order derivative. For higher order derivatives, the similar operation is progressively repeated. The main disadvantage of derivative methods is that interpretation of the spectra is made more complicated.

Figure 6 shows the impact of these filtering techniques on 50 soybean spectra collected on a Foss Infratec 1241 (spectral range from 850 nm to 1048 nm at 2 nm increments, FOSS North America, Eden Prairie, MN, USA).

2.2.3.2.2. Normalization

The second type of spectral treatment is normalization. The goal of normalization is to limit scaling differences that arise from pathlength differences (mostly present in whole sample analysis), from scattering (travel of the light in the sample that can be different depending on the size of the particles), and from hardware variations. Normalization methods "attempt to correct for these kinds of effects by identifying some aspects of each sample which should

be essentially constant from one sample to the next, and correcting the scaling of all variables based on this characteristic." (Wise et al, 2006).

The main normalization techniques are:

(1) Normalize – It is a method that enhances the similarity among samples. The most common normalizations techniques transform samples to unit length or unit area.

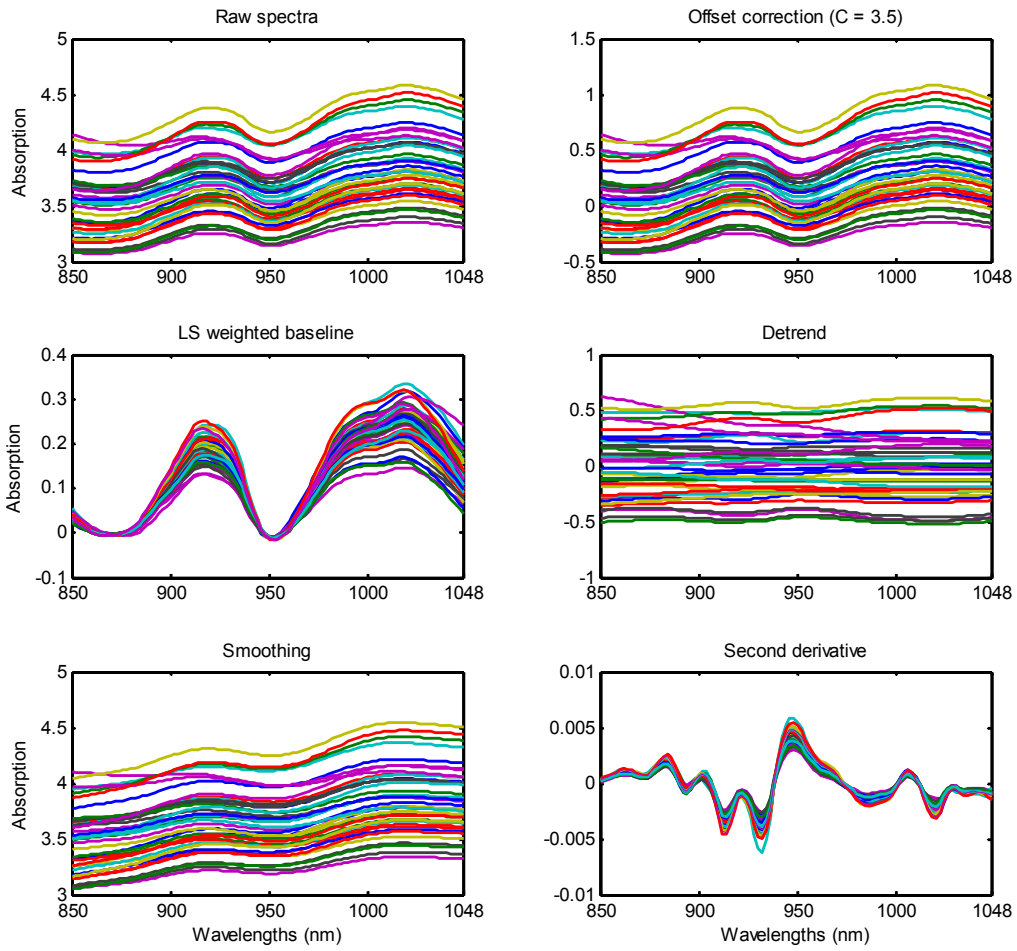


Figure 6: Spectral processing using noise removal and filtering techniques.

(2) Standard Normal Variate (SNV) – It is a weighted normalization where each variable of a sample is subtracted by the sample mean and divided by the sample standard deviation across the spectral points. The formula is as follows:

$$X_{ij} = \frac{x_{ij} - \overline{x}_i}{\sqrt{\frac{\sum_{j=1}^{n}\left(x_{ij} - \overline{x}_i\right)^2}{n-1}}} \quad (5)$$

where $\mathbf{X_{ij}}$ is the corrected absorption value for the i^{th} sample and the j^{th} variable, $\mathbf{x_{ij}}$, the raw absorption value, $\overline{x}_i$, the mean of the sample *I*, and ***n*** the number of variable for each sample. A complete description of SNV was provided by Barnes et al (1989).

(3) Multiplicative Scatter Correction (MSC) – It works by regressing a spectrum against a reference spectrum (typically the average spectrum of the considered dataset) and each variable of the measured spectrum is corrected by the slope and the offset of the fit.

Figure 7 shows the impact of normalization techniques on the 50 raw soybean spectra. The three methods of normalization were able to isolate one spectrum from the 49 others while filtering techniques did not provide this information (except 2nd derivative at one peak (at 940 nm) on figure 6). This spectrum would certainly be removed as an outlier in a calibration process.

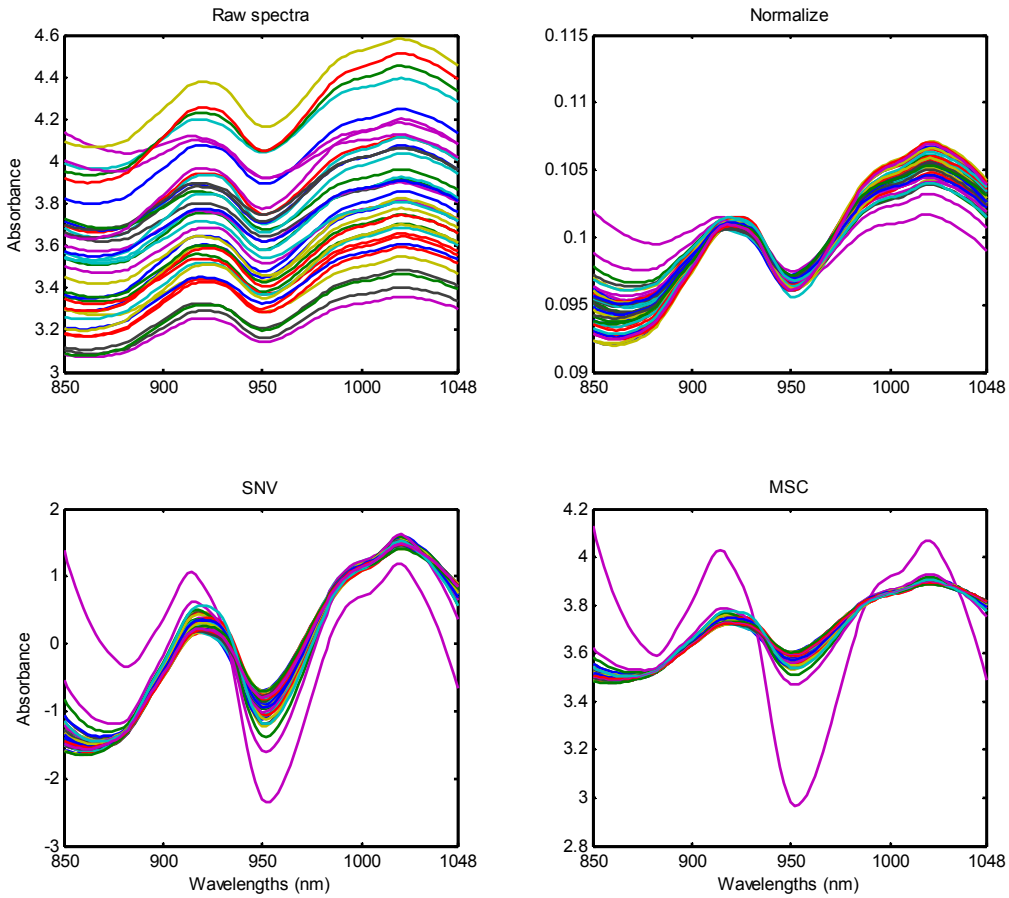


Figure 7: Spectral processing using sample normalization techniques. Unit area method was used to normalize the data.

2.2.3.2.3. *Variable scaling*

The third type of spectral processing technique concerns the scaling of variables. Most regression techniques assume that the magnitude of the measurement is proportional to its importance. This is not always true, especially in NIRS where baseline or noise effects can interfere with the sample absorption. Two techniques are often used:

(1) Autoscaling – It is the equivalent of SNV but instead of averaging and measuring the standard deviation across samples, it is done across variables. The calculation is:

$$X_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2}{n-1}}} \quad (6)$$

where $\mathbf{X_{ij}}$ is the corrected absorption value for the i[th] sample and the j[th] variable, $\mathbf{x_{ij}}$ the raw absorption value, $\overline{x}_j$ the mean of the variable *j*, and ***n*** the number of variables for each sample.

(2) Mean centering – It consists of subtracting the overall sample spectral mean $\overline{x}_i$ from each sample $\mathbf{x_{ij}}$.

$$X_{ij} = x_{ij} - \overline{x}_i \quad (7)$$

Figure 8 presents the impact of scaling methods on the 50 soybean spectra.

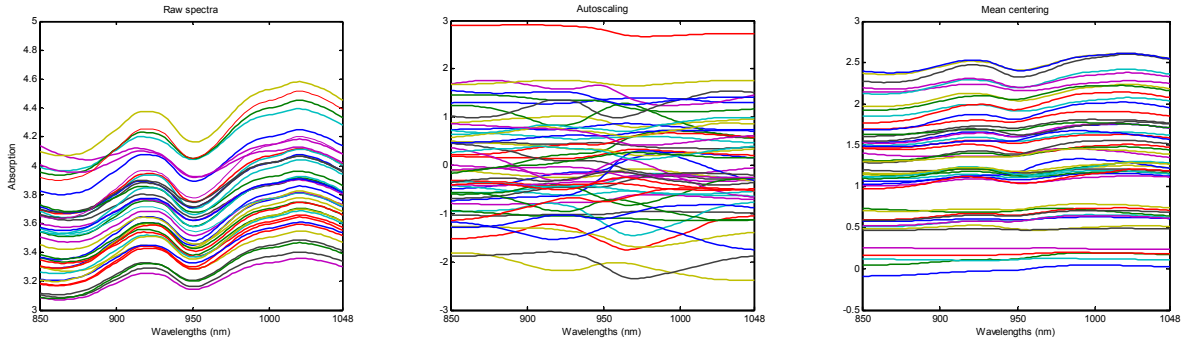


Figure 8: Spectral processing using variable scaling techniques.

The use of spectral processing methods prior to the calibration development has been, and is still, a large area of research in NIRS. Often, several methods are coupled such as derivative + SNV or MSC + autoscaling. Fernandez-Cabanas et al (2006) published an extensive study on the use of preprocessing in feedstuffs. The authors were not able to determine clear conclusions on the ability of specific processing techniques to improve results. Igne et al (2006, 2007b) showed that the use of preprocessing techniques is instruments and parameter dependent. Their use is an iterative process to find the right method and the right settings to improve the regression model. However, a calibration is rarely developed without any prior treatment of the data.

A more complete description of spectral pretreatment techniques discussed above can be found in Næs et al (2002) and Wise et al (2006).

2.2.3.3. Orthogonal filtering methods

In addition to these classical techniques, new methods have been developed to remove, at the multivariate level, the noise and other interferences. Orthogonal methods aim to remove the variance that is orthogonal (not related) to the compound to predict from the spectral data (raw or preprocessed).

Orthogonal methods can be used for both improving the signal to noise ratio and the transferability of calibration models. The theory behind these methods will be discussed in depth in section 2.2.6 since they present a large potential for removing instrumental differences. Nevertheless, it is important to note that orthogonal signal correction (OSC), orthogonal partial least squares regression (O-PLS), and dynamic orthogonal projection (DOP) can be used to remove unwanted signal and thus behave as preprocessing methods (Wold et al, 1998; Trygg et al, 2002; Zeaiter et al, 2006).

2.2.3.4. Signal transforms

A challenge in the early days of NIRS was to deal with a large amount of data with a very limited storage system and low calculation power. Signal compression methods have been developed, based on the conversion of the spectra (in the spatial/time domain) to periodigrams in the Fourier domain and approximation and detail compounds in the wavelet domain. The idea behind these methods being that some frequency components could be removed without altering the information contained in the original signal.

2.2.3.4.1. *Fourier transform*

The Fourier transform defines a relationship between a signal and its representation in the frequency domain. During the transform, no information is created or lost; the original signal can be recovered from its Fourier transform (Jackson, 2001). The theory behind the Fourier transform is that any periodic function can be represented by a Fourier series of sine and cosine functions of different amplitudes and related frequencies (McClure, 2008). A discrete spectrum in the wavelength domain could then be represented as follows:

$$F(\lambda) = a_0 + \sum_{k=1}^{M-1} a_k \cos\left(\frac{2\pi k\lambda}{N}\right) + \sum_{k=1}^{M-1} b_k \sin\left(\frac{2\pi k\lambda}{N}\right) + a_M \cos(\pi\lambda) \tag{8}$$

where a_0 is the average value of the spectrum, N (λ =1,2, …, N) represents the wavelength index and M (k = 1, 2, …, M) corresponds to the Fourier index. The array a_0, a_1, b_1, a_2, b_2, … , a_{M-2}, b_{M-2}, of Fourier coefficients is ordered, by pair, in increasing frequency to form a periodigram. Fourier coefficients represent only frequency components and do not possess any localization (spatial) information. An inverse Fourier transform can convert a periodigram to its corresponding spectrum.

Fourier coefficients are complex numbers and can be expressed in terms of magnitude and phase. The magnitude of a Fourier coefficient determines the amplitude of the contributing term. The relationship is as follows:

$$X = |X| \times e^{j\theta} \tag{9}$$

where $|X|$ is the magnitude and θ the phase angle. Also, because **X** is complex, it can be written in terms of real and imaginary part X_r and X_i respectively.

Relationships between $|X|$, θ, X_r, and X_i are as follows:

$$|X| = \sqrt{X_r^2 + X_i^2} \tag{10}$$

$$\theta = \tan^{-1}\left(\frac{X_i}{X_r}\right) \tag{11}$$

$$X_r = |X|\cos(\theta) \tag{12}$$

$$X_i = |X|\sin(\theta) \tag{13}$$

Figure 9 presents the periodigram of a near infrared spectrum, as well as its magnitude, phase angle, real and imaginary components. It is important to notice that the periodigram (figure 9.B) is symmetrical. For a spectrum of 100 data points, the first value of the periodigram represents the mean of the original spectrum (0 here). The next 50 are Fourier coefficients. Then, the end of the periodigram is the image of the 49 first Fourier coefficients with the mean omitted. This means that only 51 data points need to be stored for a complete reconstruction of the original spectrum from its Fourier transform. It is possible to use Fourier coefficients to develop calibration models (McClure, 2008).

2.2.3.4.2. Wavelet transform

The wavelet transform is similar, in many aspects, to the Fourier transform because the original spectrum is decomposed using wavelets, waveforms of specific shapes. A wavelet is a function, Ψ, with a zero average and normalized ($\|\Psi\| = 1$). The Fourier series can be considered as a specific family of wavelets using sine and cosine functions.

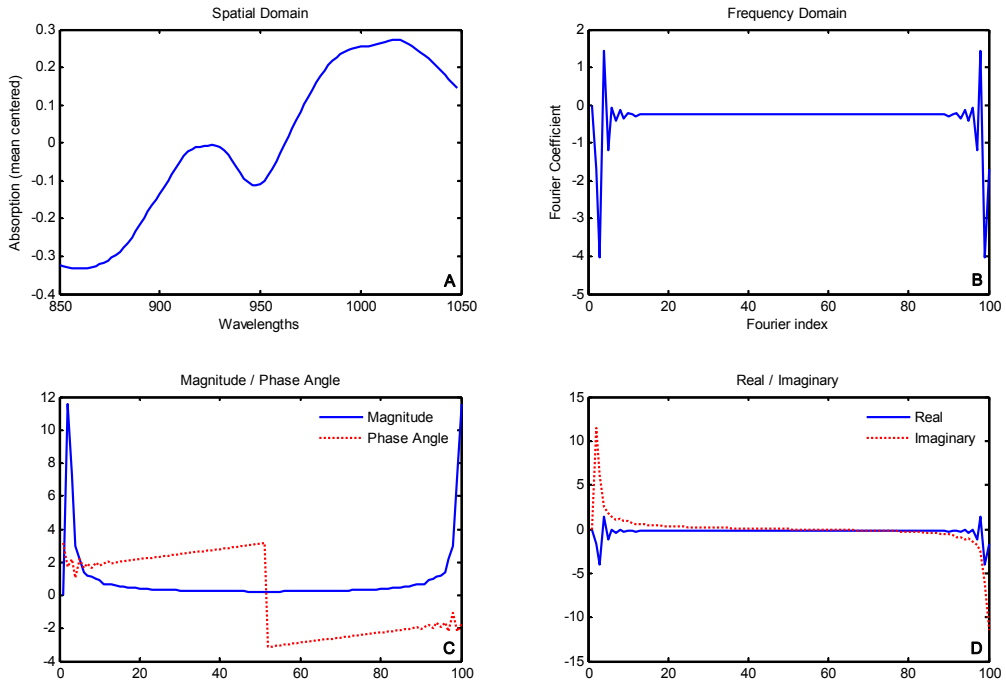


Figure 9: A spectrum and its Fourier transform. **A**: mean centered near infrared spectrum; **B**: Periodigram; **C**: Magnitude and Phase Angle components; **D**: Real and Imaginary components.

The main difference between Fourier and wavelets is that while Fourier coefficients refer only to the frequency, wavelet coefficients have both frequency and position information. Wavelet coefficients can be expressed as follows:

$$c_{j,k} = \int_{-\infty}^{+\infty} f(t)\psi_{j,k}(t)\,dx, \qquad j,k \in \mathbb{Z} \tag{14}$$

where $\Psi_{j,k}$ represents a generated function from a mother wavelet through scaling (j) and translation (k) and f(t), the original spectrum. Thus, a signal s can be decomposed into its waveform at a different scaling level j following the equation:

$$s_j = \sum_k c_{j,k}\psi_{j,k} \tag{15}$$

More practically, the wavelet transform consists of applying a low-pass and a high pass filter, respectively, called scaling filter (**H**) and wavelet filter (**G**) to the original signal. These filters are orthogonal, specific to each wavelet form and their outputs are known as

approximation and detail components respectively. Note that the scaling filter is a smoothing filter. At scaling level 1, the original signal **S** can be written as

$$\mathbf{A^1 = SH} \text{ and } \mathbf{D^1 = SG} \tag{16}$$

where **A** is the approximation and **D** the detail component. For **S** of size m x n, $\mathbf{A^1}$ and $\mathbf{D^1}$ are both of size m x (n/2). The decomposition can be carried on to higher scaling values by applying to the approximation component these same **H** and **G** filters. This is known as the Mallat's pyramid algorithm (Mallat, 1999). The decomposition process is presented on figure 10.

It is also possible to recompose approximation and detail components to get back to the original spectrum. Equation 17 presents the decomposition/reconstruction relationship for a 4th level decomposition.

$$\mathbf{S = A^4H^T + D^4G^T + D^3G^T + D^2G^T + D^1G^T} \tag{17}$$

where $\mathbf{A^4}$, $\mathbf{D^4}$, $\mathbf{D^3}$, $\mathbf{D^2}$, $\mathbf{D^1}$ are the approximations at scaling level 4 and detail at scaling levels 4, 3, 2, 1 respectively. Note that $\mathbf{A^4H^T + D^4G^T = A^3H^T}$.

One of the advantages of such decomposition is that it is possible to zoom into specific regions of the time-frequency domain, and extract waveforms from a given signal which belong to a specific scaling (decomposition level) and different time intervals. Thus, it is possible to study a specific type of frequency occurring at a specific region of the spectrum.

Various types of wavelet forms have been developed. Fear et al, 2003 studied the compression ability of wavelet transforms on near infrared spectra and determined that Daubechies 4 was the most adapted because of its smoothing abilities. Figure 11 presents the Daubechies family as well as the Coilflets and the symlets families, three very common wavelets for signal filtering in NIRS.

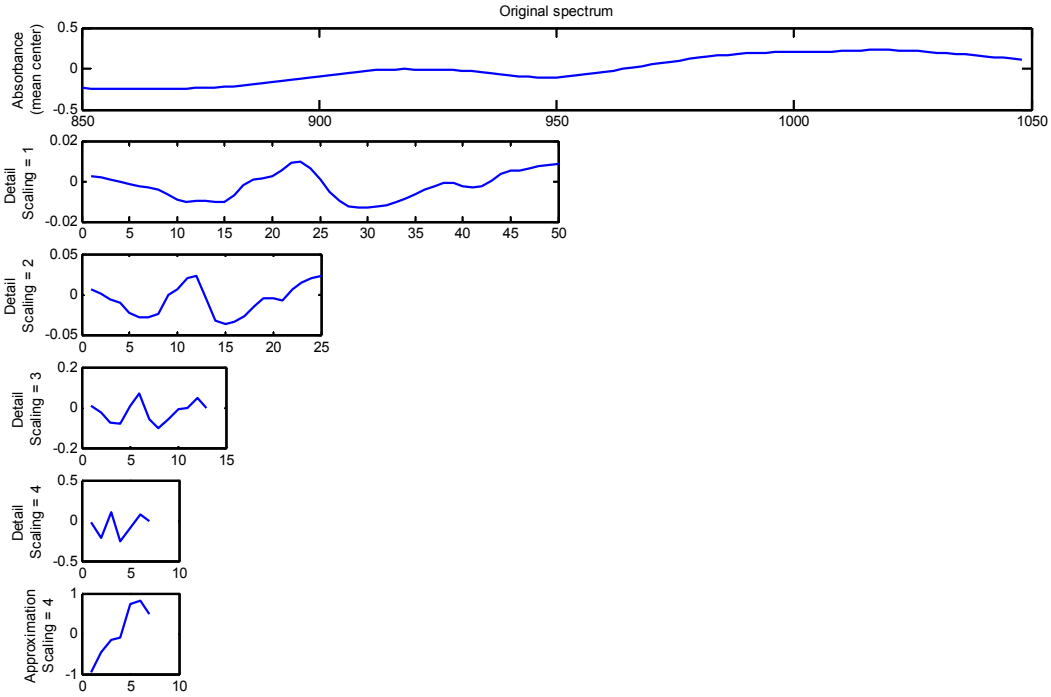


Figure 10: A spectrum and its wavelet decomposition using Daubechies 1.

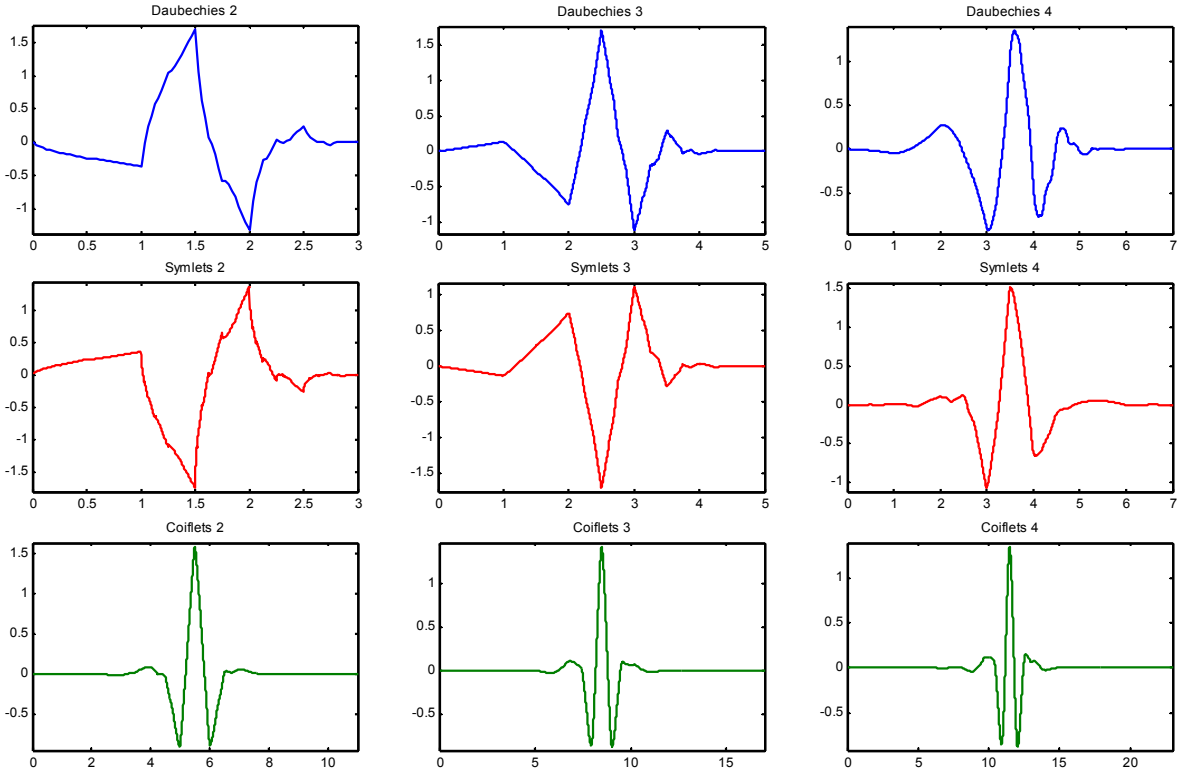


Figure 11: Most common wavelets used in signal processing.

2.2.3.4.3. *Fourier and wavelet transforms in NIRS*

As mentioned at the beginning of this section, frequency based methods have been primarily used for signal compression. Using Fourier transform, we have seen that only 51% of the data need to be kept to allow a complete reconstruction of the spectrum. Further compression can be made by implementing a low-pass filter and setting to zero high frequency components, susceptible to represent noise. With wavelets, a similar process can be done. Only parts representing information is kept and detail coefficients can be set to zero. This has the effect to keep only the element of interest and to remove the noise in a time-frequency manner. Fear et al (2003) showed that wavelet transform was generally the most efficient method.

Methods dedicated to remove noise have also been developed. Alsberg et al (1997) used moving averages on the frequency components of pure infrared spectra. McClure (2008) mentions the use of boxcar, polynomial, and Fourier smoothing to reduce noise. However, there is no description of the effect on the prediction ability of calibration models developed on frequency smoothed spectra.

2.2.3.5. *Calibration development*

As mentioned by Kovalenko (2006), the "development of calibration model is the most important and complicated step […]. Due to its complexity and comparatively large number of available methods, modeling is as much an art as a science". In the calibration process, the chemometrician's goal is to develop a relationship between the spectral data (multiple independent variables) and the reference values (dependent variable). For linear models, this relationship takes the form of a set of regression coefficients while for non-linear methods, such as artificial neural networks and least squares support vector machines, the relationship becomes a weighted structure.

Spectroscopists have adapted techniques developed in various fields. A popular example concerns partial least squares regression that was originally developed by Herman Wolds in the late 1960s to solve econometrics problems. It is now the basic technique for

full-spectrum (data presenting a high collinearity) regression situations. Currently in NIRS, three linear methods (multiple linear regression, principal component regression, and partial least squares regression) and two non-linear methods (artificial neural networks and least squares support vector machines) are used extensively. Næs et al (2002) and Wise et al (2006) provide a very good guide to calibration methods. The next subsections summarize these methods as well as more simple methods from which they have been developed.

2.2.3.5.1. Classical least squares

The assumption under the classical least squares[1] model (CLS) is that a measured spectrum is a sum of pure component spectra weighted by the concentration of the analytes. The model can be written as follows

$$\mathbf{x} = \mathbf{pS} + \mathbf{e} \tag{18}$$

where **x**, the measured spectrum, is function of **p** the weights (concentration for each element) and **S** a matrix of pure spectra plus an error term **e**. In a quantitative analysis, users are interested to know in which proportion each component is present in **x** and thus solve

$$\mathbf{p} = \mathbf{xS}^{+} \tag{19}$$

with $\mathbf{S}^{+}$ the pseudo inverse of **S**, defined such as $\mathbf{S}^{+} = \left(\mathbf{S}^{\mathbf{T}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathbf{T}}$. Thus, to determine **p**, the pure spectrum of each spectrally active component must be known or estimated. This is done by a calibration process involving, separately, each spectrally active component

$$\mathbf{S}_{\mathbf{est}} = \left(\mathbf{C}^{\mathbf{T}}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathbf{T}}\mathbf{x} \tag{20}$$

where an estimate of pure spectra $\mathbf{S}_{\mathbf{est}}$ is function of the concentration of the active component and **x** a matrix of measured spectra.

---

[1] The least squares loss function is an optimization system that tries to minimize the sum of the residuals (sum of differences between predicted and actual values) squared. This optimization function is particularly sensitive to outliers. (Abdi H., 2003).

The disadvantage of CLS lies in the difficulty to determine the pure component spectra of all spectrally active molecules for complex mixtures. The impact of noise and other hardware interferences can make this technique unstable.

An alternative to CLS is the inverse least squares method (ILS) that assumes that a regression vector can be used to determine the relationship between the y-matrix (dependent variable or the reference values) and the **X**-matrix (independent variable or optical data).

2.2.3.5.2. Inverse least squares

In the ILS situation, the relationship between the spectral data and the reference value is obtained by a set of regression coefficient **b** such that

$$y = \mathbf{xb} + \mathbf{e} \tag{21}$$

with **x** the spectrum and **y** its associated content in the parameter of interest; **e** being the error term. Those regression coefficients or beta coefficients are estimated as follows

$$\mathbf{b} = \mathbf{X}^{+} y \tag{22}$$

with the pseudo inverse $\mathbf{X}^{+}$ calculated in the most common cases with the multiple linear regression (MLR) method:

$$\mathbf{X}^{+} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \tag{23}$$

MLR is used in NIRS in conditions where the collinearity among independent variables is limited and where the number of observations is larger than the number of optical variables. As a consequence, MLR fails to develop calibration models for instruments collecting wavelength absorptions at a very short interpolation because variables are highly correlated. Also, in full-spectrum situations, the number of recorded wavelength can vary from 100 (Foss Infratec 1241) to 872 (Foss NIRSystem 6500, Laurel, MD, USA) or 1750 (ASD LabspecPro, Analytical Spectral Devices, Boulder, CO, USA) and often, the number of samples collected for calibration is limited. Williams (2001) noted that MLR is subject to overfitting when the number of independent variable is too limited.

However, MLR has been successfully applied to discrete situations (filter instruments and situations where wavelengths were preselected (Fox et al, 2002)) and where the optical data were already treated to remove the collinearity (Pienado et al, 2006).

To cope with these various issues (collinearity among variables, sparse samples, tedious choice of the number of variable to include), chemometricians have been using data compression techniques followed by regression methods to first remove the collinearity and compress the data to a few relevant variables that explain most of the variance of the **X**-matrix and secondly perform a regression on those transformed data.

2.2.3.5.3. Principal component regression

Esbensen (2004) defines principal component regression (PCR) as follows: "PCR can […] be thought of as a *two-step procedure*: first a PCA is used to transform X. The resulting T-matrix is then plugged directly into the MLR model […], now giving

$$y = \mathbf{Tb} + \mathbf{f} \tag{24}$$

instead of $y = \mathbf{Xb} + \mathbf{f}$".

The first step of PCR is based on the decomposition of the original **X**-matrix in principal components by a principal component analysis (PCA). PCA is an exploratory statistical method that aims to reduce the dimensionality of the data to reduce the noise as well as identifying new meaningful underlying variables. In a PCA analysis, eigenvectors and eigenvalues are calculated for each sample. The first eigenvector is a vector that defines the direction of the maximum variance in the data. The first principal component (PC) is a linear projection defined by the first eigenvector. Eigenvalues are the orthogonal distances from each sample to the eigenvector (they give the variance of the data in the direction of the eigenvector). The second principal component is orthogonal to the first one and represents the direction of the second maximum variance. Scores **T** are the projections of the observations in the new space. Loadings **P** give information about the relationship between the original variables and the system of principal components. "Loadings construct the

direction of each PC relative to the original co-ordinate system since the PCs can be viewed as a linear combination of the original unit vectors" (Esbensen, 2004). Thus, the **X**-matrix can be written as follows

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}^T_1 + \mathbf{t}_2\mathbf{p}^T_2 + ... + \mathbf{t}_k\mathbf{p}^T_k + ... + \mathbf{t}_r\mathbf{p}^T_r \quad (25)$$

where **r** is the rank of the matrix. The maximum number of PCs (value of **r**) is either *m*-1 or *n*-1, whichever is the smaller with *m* and *n* the number of observation and the number of variable of the **X**-matrix respectively.

Often, the PCA model is truncated to **k** PCs because after the few first PCs, the variance brought by additional PCs is often overcome by noise. The final form of **X** can be written as follows

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}^T_1 + \mathbf{t}_2\mathbf{p}^T_2 + ... + \mathbf{t}_k\mathbf{p}^T_k + \mathbf{E} \quad (26)$$

PCA then allows a compression of the data. The number of PCs included in a calibration model obviously depends not only on the matrix of interest (a simple mixture will require less PCs than a raw foodstuff) but also on the noise of the instrument. The number of PC to include in a calibration is part of the "art" mentioned by Kovalenko. Validation methods are often performed to determine the optimal number of PCs; the one that will maximize the precision and accuracy as well as limit overfitting. These methods will be discussed later.

At the end of the PCA, a matrix of scores and a matrix of loadings are obtained. PCs being orthogonal to each other, the scores do not present the collinearity problem that MLR suffered from when applied to full-spectrum situations. The new data are now ready to be plugged in a regression algorithm. This is what Esbensen meant in equation 24 where **T** signified the score matrix of **X**. When developing the MLR model, the pseudo inverse $\mathbf{X}^+$ in the system $\mathbf{b} = \mathbf{X}^+ y$ can be written as follows

$$\mathbf{X}^+ = \mathbf{P}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T \quad (27)$$

where **P** is the loading matrix and **T** the score matrix.

PCR, as well as MLR, is included in every chemometrics suite. This is a very powerful technique that has been at the origin of the use of wide spectral ranges for the

development of various applications in NIRS (Sun, 1996; Park et al, 2001). However, PCR remains tedious to use because of the possible error in choosing the appropriate number of PCs. Partial least squares (PLS) regression was developed to simplify this dilemma.

2.2.3.5.4. Partial least squares regression

Wise et al (2006) present the PLS regression as follows: "Partial Least Squares (PLS) regression […] is related to both PCR and MLR, and can be thought of as occupying a middle ground between them. PCR finds factors that capture the greatest amount of variance in the predictor (X) variables […]. MLR seeks to find a single factor that best correlates predictor (X) variables with predicted (Y) variables […]. PLS attempts to find factors which both capture variance *and* achieve correlation."

PLS, similarly to PCR, decomposes (compress) the original **X**-matrix into principal components. However, PLS determines its principal components by maximizing the covariance between the y-matrix and all possible functions of the **X**-matrix. Thus, PLS principal components are more directly related to the variability in the y-matrix than PCA principal components are. The **X**-matrix is decomposed in scores (**T**), loadings (**P**), and an additional set of vectors called weights (**W**). Weights are computed from **X** directly and are necessary to calculate **T**; they are required to keep scores vectors orthogonal to each other. In PCA, $\mathbf{T} = \mathbf{XP}$ while in PLS $\mathbf{T} = \mathbf{XW(P^TW)^{-1}}$. As a consequence, the pseudo inverse $\mathbf{X}^+$ in the system $\mathbf{b} = \mathbf{X}^+ y$ can be written as follows

$$\mathbf{X}^+ = \mathbf{W}\left(\mathbf{P}^\mathbf{T}\mathbf{T}\right)^{-1}\left(\mathbf{T}^\mathbf{T}\mathbf{T}\right)^{-1}\mathbf{T}^\mathbf{T} \tag{28}$$

where **P** is the loading matrix, **T** the score matrix, and **W** the weight matrix. More theoretical explanations of PLS can be found in Wold et al (1984), Geladi and Kowalski (1986), Lorber et al (1987), and Martens and Næs (1989).

The main advantage of PLS over PCR is the reduction of the complexity of the models. Table 2 provides an example of calibrations developed with both methods and shows that while PCR needed eight PCs to explain 90% of the variability of **X**, PLS required only four factors. This has an impact on the future predictive ability of models since the 4-

factor model will be more robust (less influenced by noise). While with 94% of the variability, PLS could achieve the 90% predictive level, PCR required 99.7% of the spectral variance (Boysworth and Booksk, 2001). It is necessary to validate PLS models to ensure an appropriate selection of the number of PCs.

MLR, PCR, and PLS are the three main linear methods used in NIRS. However, these techniques have difficulties in situations where the relationship between the independent variables and the dependent variable is non-linear or when the distribution of the residual is not normal. Chemometricians have been using techniques such as artificial neural networks (ANN) and support vector machines (SVM) to deal with non-linearity and outliers.

Table 2: PCR and PLS comparison on the ability to extract data from X-matrix (Boysworth and Booksk, 2001).

Factor	PCR		PLS	
	Cumulative % spectral variance	Cumulative % predictive variance	Cumulative % spectral variance	Cumulative % predictive variance
1	88.21	3.53	82.54	13.96
2	94.78	50.60	94.73	57.30
3	96.94	50.61	95.58	83.58
4	97.92	60.86	96.15	92.73
5	98.68	62.93	97.94	93.58
6	99.14	71.76	98.72	94.40
7	99.45	72.10	99.33	94.84
8	99.70	93.15	99.44	96.49
9	99.81	93.16	99.66	96.83
10	99.87	93.83	99.78	97.34

2.2.3.5.5. Artificial neural networks

Artificial neural networks have brought a whole new dimension to the development of near infrared calibrations. While unsuccessful on small datasets, ANNs have shown advantages over linear methods when dealing with large databases. Numerous publications show cases where ANNs provided more accurate and precise results than linear techniques (Wu et al, 1996; Dou et al, 2005; Kovalenko et al, 2006). Furthermore, some commercial prediction models are now ANN-based (Foss Infratecs).

As stated in its name, ANN mimics the functioning of biological neural networks by being constituted of individual neurons positioned on interconnected layers. A typical ANN has an input layer, one or more hidden layer(s), and an output layer. The number of neuron on the input and the hidden layer(s) can vary. Figure 12 presents an example of three-layer structure. Kecman (2001) provides a very good description of the ANN algorithm.

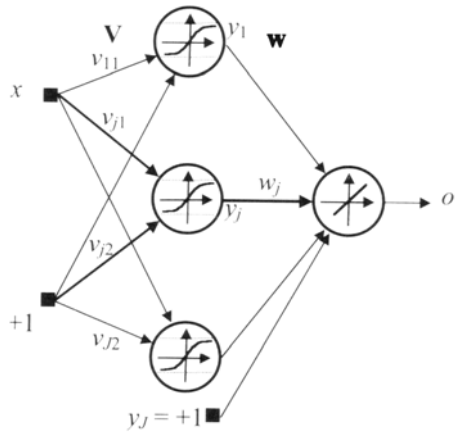


Figure 12: ANN network structure: 2 neurons on the input layer, 3 neurons on the hidden layer, and a single output (Kecman, 2001).

An activation function (or transfer function) is associated with each neuron of the hidden and output layers. In a biological neural structure, if the input intensity does not reach a threshold, no information is passed along. In an artificial neuron, an activation function is responsible for the transmitted value. Several functions exist, but in NIRS, the most used are the log-sigmoid, the tan-sigmoid, and the linear transfer function. Inputs are typically PCA or PLS scores (orthogonal vectors). The number of input neurons (number of PCs) has to be optimized for the application. This value is typically higher than the number of PCs used in linear models. Inputs are multiplied by a randomly defined weight and sent to all neurons of the first hidden layer. Inputs are summed and applied to the transfer function. In a three layer network, outputs of the hidden layer are multiplied by a weight and sent to the output layer where they are summed and applied to the transfer function to obtain predictions. These predictions are compared with reference values and a calibration error is determined. If the value is greater than a threshold set by the user, an optimization

process such as the backpropagation algorithm starts, consisting in adapting the weights to reduce the calibration error.

In addition to the number of input neurons, the internal structure of the network, and the activation function, the user has many other parameters to set: algorithm to set the original weights, algorithm to adapt them. Also, depending on the learning rate (a small learning rate will modify weights moderately), an ANN can require a large number of epochs to converge. An epoch is a cycle of backpropagation and adaptation of the weight, from the output layer to itself. The momentum is another important parameter to set. It acts like a low-pass filter and allows the network to not stop its training on a local minimum. The training of a network can stop when it reaches the user's set performance or when the error does not decrease during a defined number of epochs. Thus, a low momentum increases the risks to find a local minimum, but a too large momentum can cause the network to never converge. The training will then stop when the maximum number of epoch is reached, or when the allowed training time is exceeded.

In terms of structure, most of the functions can be emulated with just one hidden layer. However, the number of neurons can vary. A good method consists of starting with many and prune (remove branches of the network with weights equal or close to zero) to simplify the network.

Regarding the number of samples needed, two conditions exist regarding the structure of the network:

Necessary condition:

$$\text{(Number of samples)} > \frac{W}{(1-a)}$$

Sufficient condition:

$$\text{(Number of samples)} \geq \log\left(\frac{N}{(1-a)}\right) \times \frac{W}{(1-a)}$$

with **W** the number of weights, **a** the desired accuracy and ***N*** the number of neurons.

ANN is particularly subject to overfitting. A good validation strategy must be used. ANN performs poorly when trying to predict values outside of the training range, while linear methods can extrapolate quite well. Also, it is almost impossible to develop the same model twice because of the random assignment of initial weights. This is a major issue of traceability because often, it is not possible to recreate a model and for heavily regulated environments such as pharmaceutical industries, ANN cannot be used. Also, no actual chemometrics suite supports ANN and it has to be used offline.

In the same logic of learning, least square support vector machines were developed to establish the relationship between independent and dependent variables presenting a non linear relation in the situation of limited data availability.

2.2.3.5.6. Least squares support vector machines

Support vector machines were originally developed for binary classification situations (Vapnik, 1995). The idea behind SVM is to determine samples that define the most appropriate cluster limits (support vectors) as well as reduce the misclassification rate. In regression situations, SVR (support vector regression) and LS-SVM (least squares support vector machines) try to find the best fit of the data by limiting the number of samples outside a range error (or insensitivity zone) set by the user (see figure 13).

The difference between ANN and SVM is that while ANN is focusing only on the closeness of the model to the data (which implies risks of overfitting and poor extrapolation), SVM adds a *structural risk minimization* parameter which, in practice, signifies a unique solution to the problem (called capacity of the machine). This means a significant difference in treating the error between the two techniques. In ANN, the confidence interval on the prediction is fixed and the network learns how to minimize the training error. In SVM, the training error is fixed (equal to zero or an acceptable level) and the confidence interval is minimized.

In terms of algorithms, SVR and LS-SVM consist in solving the following optimization system:

Minimize:

$$\frac{1}{2}W^T W + C\sum_{i=1}^{n}\left(\zeta_i + \zeta_i^*\right)$$

Constraints:

$$y_i - \hat{y}_i \le \varepsilon + \zeta_i$$
$$\hat{y}_i - y_i \le \varepsilon + \zeta_i^*$$
$$\zeta i, \zeta_i^* \ge 0$$

with **W** the weights, **C** the penalty factor, **ε** the insensitivity zone, and ζ defined such as $\zeta = |y - f(x,w)| - \varepsilon$ for above the insensitivity zone and ζ^* bellow. This optimization problem can be solved by a quadratic approach with constraints that ends in forming and solving a primal and dual Lagrangian (SVR) or by solving a set of linear equations (LS-SVM). In addition, the kernel "trick" allows a modification of the space to make the relationships between y and **X** linear. The major kernel functions are polynomials (nth degrees) and radial basis functions (RBF).

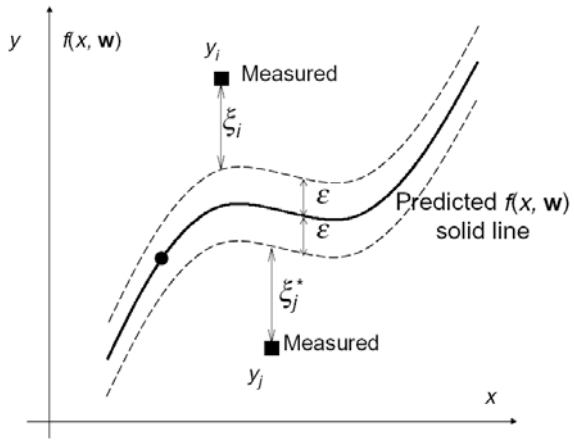


Figure 13: In SVR and LS-SVM, the minimization of the error is based on the size of the insensitivity zone (2×ε) and the cost given to "outside" individuals (Kecman, 2001).

Thus, even if SVR and LS-SVM appear much more complex than ANN, their implementation is simpler because fewer parameters need to be set. The user must determine which kernel function he/she wants to use and its parameter (degree of the polynomial or shape of the bell for RBF). This has an impact on the shape of the regression

function. The regularization parameter **C** must also be set. However, numerous optimization systems have been developed to find the most appropriate set of parameters. Optimization functions such as the "gridsearch" function of the LS-SVMLab toolbox (Suykens et al, 2002) perform an exhaustive search of these elements. However, this optimization process is extremely time consuming and is exponentially dependent of the size of the dataset. An exhaustive search of the optimal parameters for a matrix with 1,000 observations and 100 wavelengths per observation can take more than a day of calculation to terminate on a Windows based computer with a processor of 2.4 GHz and 2 GB of RAM. Some authors have suggested the use of subsets from the calibration set to perform the optimization process and then validation with an independent dataset once they have been applied to the entire calibration set (Fernandez, 2006; Igne et al., 2008).

SVM is a relatively new method in NIRS even though the technique is as old as ANN. Several papers have demonstrated its good performances compared to PLS and ANN in situations of large but also reduced datasets (Chauchard et al, 2004; Kovalenko et al, 2006; Igne et al, 2008). On the contrary to ANN, LS-SVM models are reproducible and can be used in regulated environments. However, similarly to ANN, SVM is not included in any chemometrics package and it has to be employed offline, limiting its use by the NIRS community.

2.2.3.5.7. Underfitting and overfitting

An issue inherent to compression methods or based on compression methods such as PCA, PCR, PLS, and ANN is the choice of the appropriate number of principal components. Using too few PCs is called underfitting and results from not including enough variance from the calibration set into the model. With this situation, the model will be difficult to build and its predictions will be unstable. On the contrary, using too many PCs is called overfitting and results from including too much noise in the model. The model will be fit to the calibration samples, but will be unable to predict accurately new samples. A different type of error is associated with each situation, as presented on figure 14. The optimum is usually found between the two extremes. Using validation methods, such as

cross-validation, imbedded in the calibration algorithm, help to choose the best number of principal components to include.

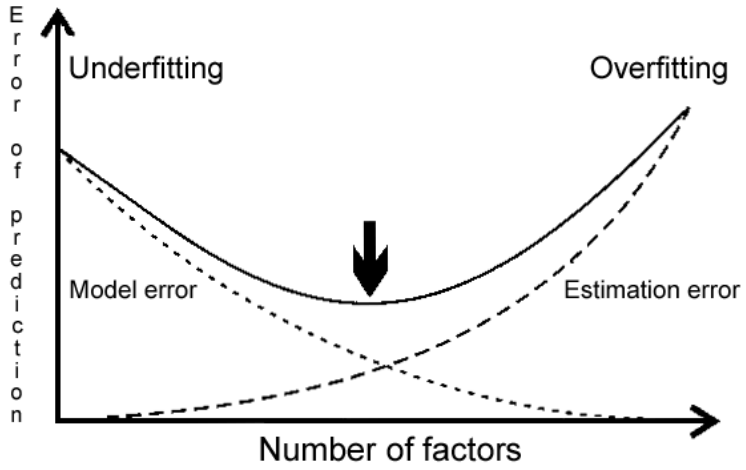


Figure 14: Representation of under and overfitting and the prediction error (Næs et al 2002).

2.2.3.6. Validation methods

The validation is the last step in the creation of a prediction model. Various forms of validations are used during the calibration process (i.e. choice of the number of PCs in PLS using a cross-validation, use of a "stop" set for ANN to limit overfitting). But a solid validation strategy must be employed to ensure the good functioning of the calibration. Three validation strategies are usually used in NIRS.

The first consists of using cross-validation methods (Næs et al, 2002). Cross-validation implies using samples of the calibration set in the validation process. Block cross-validation isolates a sequence of samples from the calibration set, develops a calibration model on the remaining samples and validates the model on left apart samples. Validation performances are stored, validation samples return in the calibration set and another set of sample is isolated for validation. The process continues until all samples in the calibration set have been part of a validation set. Validation performances are averaged and constitute

the estimated validation parameters of the calibration model developed with all the samples. The size of the block can vary. The larger the block, the more realistic the validation, but in case of sparse samples, large blocks are difficult to use. An alternative to block cross-validation is leave-one-out cross-validation. It is actually a particular case of block cross-validation where only one sample is removed from the calibration set. Cross-validation results are usually considered over-optimistic regarding the real performances of the calibration.

The second method of validation consists of isolating, prior to the development of the prediction model, a part of the calibration samples to be used as validation set. This technique is a specific case of cross-validation, except that the error of prediction comes from only one prediction event. Such a method is perfectly suitable when samples in the calibration set and samples that will be analyzed in routine analysis are similar (as in the case of synthetic products). When dealing with agricultural products and other products that change from year to year because of the climate impact and the changes in genetics brought by breeders, a more adapted validation technique has to be implemented, such as a validation based on next year or another origin samples.

Validation on another year or origin samples is the best validation method for NIRS in situations of extraneous variability brought by uncontrollable environmental and genetic changes. However, it is hard to implement because it is necessary to wait for the next crop year or new samples.

Igne et al (2007b) compared these validation methods in the development of triticale calibrations. Authors showed that while cross-validation methods give an indication of the performances of the validation, the use of part of the calibration samples for validation was closer to the real performances given by the use of next year samples.

Validation performances need to be evaluated in terms of precision and accuracy. Precision means the degree to which further measurements or calculations show the same or similar results. In NIRS, it is evaluated by the model fit (coefficient of determination (r^2)

of the comparison between actual and predicted values on the validation set) and the standard error of prediction (SEP) or standard deviation of differences corrected for bias. Statistical definitions of these terms are:

$$r^2 = \frac{\left(\sum \hat{y}y - \frac{\sum \hat{y}\sum y}{n}\right)^2}{\left(\sum \hat{y}^2 - \frac{\left(\sum \hat{y}\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)} \tag{29}$$

and

$$SEP = \sqrt{\frac{\sum (\hat{y} - y)^2 - \frac{\left(\sum \hat{y} - y\right)^2}{n}}{n-1}} \tag{30}$$

where $\hat{\boldsymbol{y}}$ are the predicted values, **y** the actual values, and ***n*** the number of samples. It is worth mentioning here that $\pm 1.96 \times SEP$ provides a 95% confidence interval on the predictions (from a normal distribution), but since the standard deviation is almost never known, the rule of thumb states that a 95% confidence interval is $\pm 2 \times SEP$ (from a t-distribution with moderate n value).

On the other hand, the accuracy is the degree of conformity of a measured or calculated quantity to its actual (true) value. The bias is used to measure accuracy. It is calculated as follows:

$$Bias = \frac{\sum (\hat{y} - y)}{n} \tag{31}$$

where $\hat{\boldsymbol{y}}$ are the predicted values, **y** the actual values, and ***n*** the number of samples.

The relationship between SEP and Bias is given by the equation 32.

$$RMSEP^2 \approx SEP^2 + Bias^2 \quad (32)$$

RMSEP is the root mean standard error of prediction and is a measure of accuracy. It is the standard deviation of the differences between observed and predicted values, not corrected for bias. Exact equality is not obtained because RMSEP uses *n* at the denominator and not *n-1* as in SEP.

A last term, specially developed for NIRS measurement, is often used to compare calibration models validated on the same validation set. It is called relative predictive determinant (RPD) and reveals the ability of a prediction model to be used for various applications (Williams, 2001). It is the ratio of the standard deviation of the actual values to the SEP. It shows how the calibration is able to predict the variation of a validation set. A scale allows an easy comparison between calibration models (see table 3).

Table 3: RPD values and their practical applications (Williams, 2001).

RPD values	Classification	Application
0.0-2.3	Very poor	Not recommended
2.4-3.0	Poor	Very rough screening
3.1-4.9	Fair	Screening
5.0-6.4	Good	Quality control
6.5-8.0	Very good	Process control
8.1+	Excellent	Any application

To complete this section on validation, it is necessary to talk about the error of the reference laboratory. NIRS is based on the measurement of a samples parameter by a first analytical method, so the error of this analytical method must be as low as possible to ensure good prediction results.

It is measured by the standard error of the laboratory (SEL) and expressed as follows:

$$SEL = \sqrt{\frac{\sum_{i=1}^{n}\left[\sum_{j=1}^{R}\left(y_{ij} - \bar{y}_i\right)^2 / \left(R-1\right)\right]}{n}} \tag{33}$$

where y_{ij} is the j[th] replicate of the i[th] sample, $\bar{y}_i$ is the reference method mean value of all the replicates of the i[th] sample, n is the number of samples, and **R** is the number of replicates.

A complete description of these validation parameters can be found in Williams (2001) and Næs et al (2002).

SEP is a function of SEL. When developing a prediction model, the error in the y-matrix is incorporated in the error of the model. In the best situation, the error of a calibration model can be as low as the error of the reference laboratory. David Hopkins (2008) presents this relation as follows

$$SEP^2 = SEP^2_{Instrument} + SEP^2_{Sample} + SEL^2 \tag{34}$$

where he decomposes SEP in three elements, the error of the instrument (repeatability), the error of the sampling (sample representativity), and the error of the laboratory. These three elements are inherent to any NIRS prediction and can be quantified, minimized but not erased.

2.2.4. Optimization methods for calibration development

At the end of the chapter on instrumentation, we discussed the fact that some instruments collect spectra with a very short interpolation (every 1 or 2 nm) over a large range while other units collect only fewer absorption values at specific wavelengths, selected for the application. Applications have been developed based on the selection of a few variables from a spectrum. For instance, Foss Analytical patented the use of four wavenumbers (1700, 1407, 1365, and 1238 cm^{-1} or 5882, 7107, 7326, and 8077 nm) to develop models for the prediction of acetone in milk (US Patent 6385549 – May 2002). Foss stated that these four wavelengths optimized the prediction of acetone in the infrared region. Variable selection is a type of optimization fairly common in NIRS. Another type of optimization is to use only the most appropriate samples in calibration, the samples that will provide the best fit for the new sample. These models are called local models.

Both optimization methods aim to reduce the complexity of the model and can be coupled. In this section, these techniques to optimize NIRS prediction models are presented.

2.2.4.1. Variable selection

2.2.4.1.1. Exhaustive search

Selecting the best variables for the prediction of a parameter is a tedious procedure. The most intuitive method consists in performing an exhaustive search. The problem of this approach is the time to perform the calculations. For an instrument collecting n wavelengths, the total number of possible combinations is $2^n - 1$. For $n = 100$, the total number of possible variable combination is 2^{100} -1 or 1.2677×10^{30}. This number is large and, even with a powerful computer, the exhaustive search of the best variable combination takes very long to perform. A type of exhaustive search is Interval PLS (Norgaard et al, 2000). Instead of selecting a single variable at a time, a group of adjacent variables are selected; they are called intervals. PLS models based on these intervals are developed and

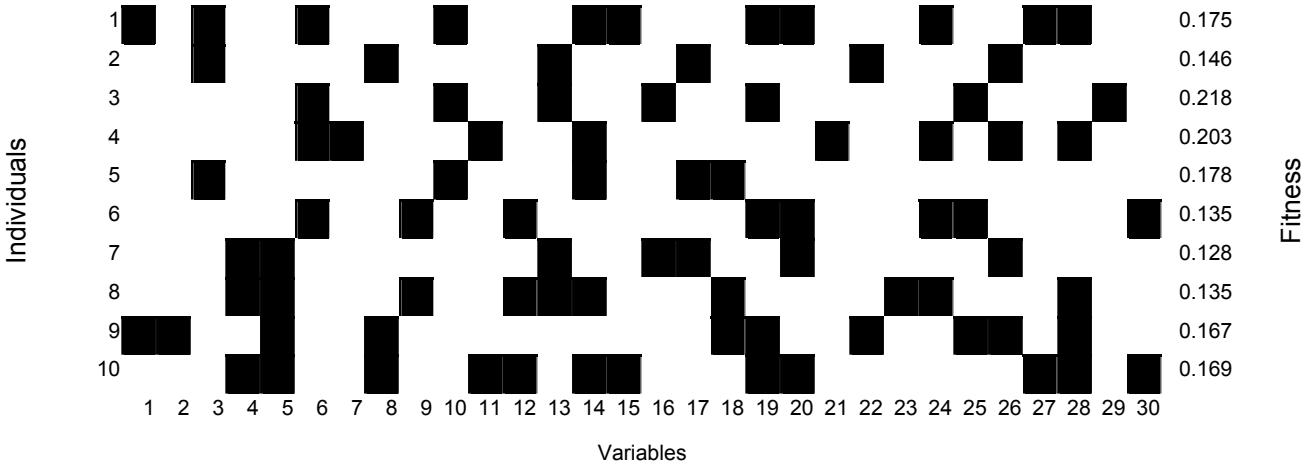
their RMSECV (Root Mean Standard Error of Cross-Validation) is calculated. RMSECV is often computed with a block cross-validation. The interval presenting the best fitness value is selected. This interval is then combined to other intervals, one at a time. New PLS models are developed and the best combination of intervals is kept. The operation is carried on until the fitness value starts increasing. A reverse-interval PLS can be implemented where the first model is developed with all intervals and intervals are removed one at a time until the lowest fitness value is reached. Note that the number of PCs to use in each model is chosen by the user at the beginning of the optimization process. The main disadvantage of interval PLS is its calculation time. When the size of intervals is low and the number of variables large, the convergence time can become very high (hours).

2.2.4.1.2. Genetic algorithms

Alternatives to exhaustive search methods have been developed based on stochastic approaches. A stochastic method is a method involving some elements of randomness. Genetic algorithms are a family of methods using stochastic search to find the best solution. Genetics algorithms copy evolution. They use gene modification methods (combination and mutation) to progressively improve the breed, in this case the fit. To use genetic algorithms for variable selection, it is necessary to randomly determine individuals, subset of consecutive (or not) variables. PLS or PCR models are developed using each individual in calibration and RMSECV is used to evaluate the fit. The number of factors to use is predetermined by the user. At the end of this first step, a table representing, for each individual, its fitness and the variable ID is obtained. The flag yes/no of each variable that determine if a specific variable is part of an individual is called a gene. Table 4 presents such map of individual/gene.

The second step discards individuals that have a fitness value (RMSECV) larger than the median. In the present example, individuals 1, 3, 4, 5, and 10 will be discarded. To "repopulate" the number of individuals, combinations are implemented and variability in the population is generated by mutations.

Table 4: Map of individuals with selected variables of a 30 wavelength spectrum (white squares represent selected variable).



For practicality, we will represent black squares with 0 and white squares with 1. Individuals 1 and 2 from the original population can then be represented as follows:

Individual 1: 010110111011100111001110110011

Individual 2: 110111101111011101111011101111

A combination consists in exchanging parts of the parent genes to create new individuals. The cross-over can be single or double. In the single cross-over, a variable is selected and everything on the left of this variable will be crossed-over to the other individual. In a double cross-over, two parts of the individual are exchanged. Thus, using Individuals 1 and 2 as parents, we would obtain with a single cross-over at variable 21, the following two new individuals:

New Individual 11: 010110111011100111001 | 011101111

New Individual 12: 110111101111011101111 | 110110011

and with a double cross-over at variables 7 and 15, we would obtain:

New Individual 11: 0101101|01111011|111001110110011
New Individual 12: 1101111|11011100|101111011101111

If only combinations were possible, offspring could not contain variables that were not selected during the initial random assignment. Mutations allow this by randomly converting a non-selected variable to a selected variable. Using individual 1 as parent, a mutation occurring at gene 10 would give:

New Individual 11: 0101101111110011100111001110110011

The generation of offspring carries on until the population of individuals is back to its original size (10 here). RMSECVs are calculated, just like after the random initial population assignment. The optimization process carries on as long as some percentages of individuals use the same variables or that the total number of generation is reached (also set by the user). At the end, the selected variables can then be used to develop a calibration model, where the number of principal components can be carefully chosen. The cross-validation process can also be replaced by an independent validation set to evaluate the real performance of the model.

Thus, through these five steps (1. random creation of individuals, 2. fitness evaluation, 3. individual removal, 4. population breeding, 5. population mutation), whose steps 2 to 5 are carried on until the algorithm stops, noisy variables (responsible for higher RMSECV) are discarded because less and less individuals are using them.

The use of genetic algorithms requires the tuning of many parameters (the number of PCs to use in the automatically generated models, the nature of combination and the rate of mutation, the number of generations allowed …) and it is also a method subject to overfitting. A consistent validation strategy should be implemented.

Genetic algorithms have been used for twenty years in NIRS, but the main issue that developers are facing is that few instrument have software able to implement them (only available in statistical suites). However, they can be easily applied into instruments because regression coefficients of non selected variables can simply be set to zero. The literature reports successful applications over full spectrum methods for soluble solid content in apple (Shi et al, 2008) and grains (Davies, 1987; Leardi et al, 1992).

The manual of the PLS_toolbox (Eigenvector Research, Wenatchee, WA, USA) is a very good reference for the theoretical and practical use of genetic algorithms.

2.2.4.1.3. Particle swarm optimization

Another stochastic approach for variable selection is particle swarm optimization (PSO). PSO represents a family of algorithms that mimic the behavior of social insects. They were introduced by James Kennedy and Russell Eberhart in 1995 and are based on three sociocognitive underpinnings: Evaluate, Compare, and Imitate. For variable selection, the use of PSO is quite similar to GAs since it is a binary combinatory problem. Continuous versions of PSO exist, but will not be discussed here. At the time this literature review was written, only one NIRS-based article has been published on the subject (Xu et al, 2004). But, PSO has been used for the optimization of ANN (to replace the back propagation algorithm) and this technique generates a lot of interest in the NIR community (Kennedy and Eberhart, 2001).

In PSO, each particle (individual in GA) – a subset of variables among the variable population – is provided with a velocity. At each generation, the fitness of each particle is compared and its velocity is updated in function of its performance at generation n and earlier, the best global particle, and the best neighboring particles. Thus, the best performing particles will be imitated by other particles and the process will carry on until the performance criterion is reached.

The pseudo code of PSO is provided below.

```
1-  Loop
2-      For i = 1 to Number of Particle
3-          If Fitness(Particle i)  < Fitness(Historical Best)
4-              For d = 1 to Dimensions
5-                  Particle_id = New Best_id
6-              Next d
7-          End If
8-          g = i
9-          For j = indexes of Neighbors
10-             If Fitness (Neighbors j) < Fitness (New Best)
11-                 Then g = j
12-             End If
13-         Next j
14-         For d = 1 to number of Dimensions
15-             v_id(t) = v_id(t-1) x W_i + φ1 x rand() x (New Best_id – …
16-                 Particle_id (t-1)) + φ2 x rand() X (Best Neighbor_gd – …
17-                 Particle_id (t-1))
18-             If ρ_id< s(v_id(t))
19-                 Then Particle_id (t) = 1
20-                 Else Particle_id (t) = 0
21-             End If
22-         Next d
23-     Nest i
24-  Until criterion
```

Step 1: Adapt based on best overall

Step 2: Find best in neighbors

Step 3: Update velocity and particles

where v_{id} represents the velocity of the particle i and dimension d at step t or t-1, W_i is the inertia weight for the particle *i*, φ1 and φ2 are learning factors whose sum is 2.0, and ρ is a random vector between 0 and 1. $v_{id}(t)$ is the parameter that determines if a dimension *d* will take the value 0 or 1 (0 meaning the variable is not selected, and 1, the variable is selected).

To obtain values between 0 and 1 the velocity is scaled using a sigmoid function.

$$s(v_{id}) = \frac{1}{1+\exp(-v_{id})} \tag{35}$$

Since PSO looks for the best performing particles in the entire population and best previous state of the particle, the convergence is often quicker than for GA. The fitness parameter is usually the RMSECV.

Since few studies have used PSO for variable selection, it is not possible to tell which method, PSO or GA, performed better for NIR spectra. More theory on PSO can be found in Kennedy and Eberhart (2001).

2.2.4.1.4. Other methods

Other optimization methods have been implemented to select variables. We can mention simulated annealing and evolutionary programming. We will not discuss these techniques, but their use is increasing in NIRS (Luke, 1994; Swierenga et al, 1998; Lu et al, 2004; Shen et al, 2004).

2.2.4.2. Local regression

Local models aim to optimize the prediction model statistics by choosing only the most appropriate samples to include in the calibration set. A model is generated each time a new sample is scanned for prediction. Tom Fearn (2001) stated that "A local calibration is one in which the equation used to predict for a given unknown sample is derived from only a subset of the available training samples, this subset having been chosen because the spectral data for the samples it contains resembles the spectral data for the unknown in some particular way".

Local models have been developed to deal with clustering issues or non linearity issues. We briefly mentioned, when presenting ANN and SVM, the non-linearity problem. A parameter can be non-linearly related to the spectral data and the use of linear regression methods will not provide the best prediction. Figure 15.A presents the relationship between a parameter, y, and a univariate **X**. A clear non-linear pattern exists. However, as displayed by figure 15.B, it is possible to select only a reduced range of data, and still be able to use linear regression methods to perform an accurate prediction over this range.

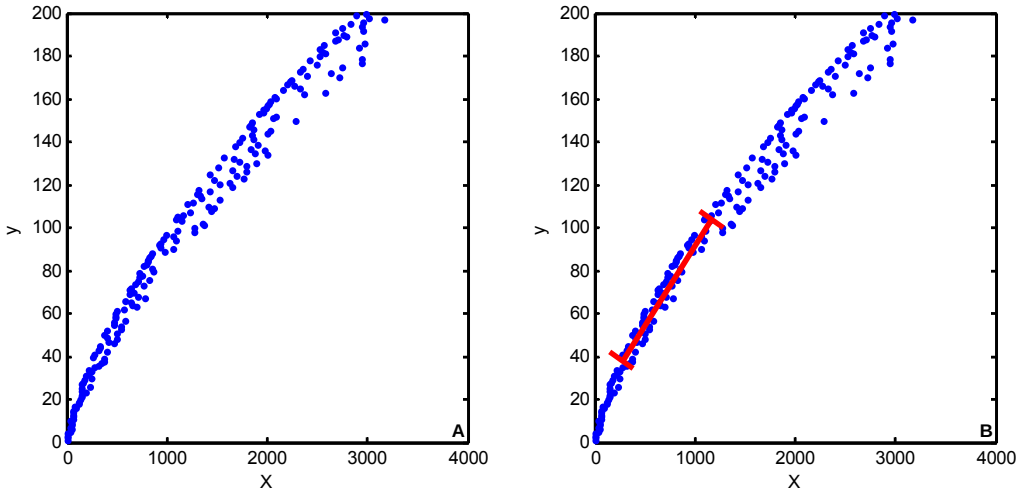


Figure 15: Non-linearity issues. Plot **A** shows a non-linear relationship between two variables X and y. Plot **B** presents a possible data range where local regression with linear regression method is possible.

The implementation of local models globally follows the same steps, no matter the algorithm used:

1. Creation of a calibration set
2. Collection of a new sample spectrum to be predicted
3. Selection of samples the "closest" or the most "similar" from the calibration set to the sample to predict
4. Prediction of the new sample based on the "local" model.

First, it is necessary to possess a large library of samples. Enough samples should be present in the neighborhood of the new samples to ensure its precise and accurate prediction. It is then necessary to select the most appropriate samples from the available

database to develop a local calibration set. Depending on the algorithm, different methods are used to find the closet samples. Here, the three most common techniques used in NIRS will be presented (CARNAC, LOCAL, and locally weighted regression).

In the method CARNAC (Davies et al, 1988; Davies et al, 2006), standing for Comparison Analysis using Restructured Near-infrared And Constituent data, the database, as well as the new sample to predict, is compressed using Fourier or wavelet transform and a similarity index is determined. This index is defined as

$$s = \frac{1}{(1 - r^2)} \tag{36}$$

where $\mathbf{r^2}$, the coefficient of determination, exposes the similarity of the new sample with those in the database. The number of samples to include can be set by either using a threshold on **s** or by selecting a fixed number of samples, among the closest. The algorithm LOCAL (Shenk et al, 1997) performs similarly by determining the correlation coefficient (r) between the spectra in the database and the new spectrum (US Patent 5798526 – August 1998). Spectra can be raw or pretreated, but no compression method is used. Similarly, the number of samples to keep is determined by a threshold or a fixed number of samples to include.

The locally weighted regression (LWR) algorithm (Cleveland et al, 1988, Næs et al, 1990) uses a different approach to find the closest spectra. A PCA compresses the database and the sample to predict. The closest samples are determined by calculating Euclidian or Mahalanobis distances between scores of the new sample and scores of the calibration set. A modification to LWR (LWRY) introduced the possibility to add information from the Y-matrix in the distance calculation (Chang et al., 2001).

The Mahalanobis distance was replaced with a more complex distance expression:

$$D_i = \alpha Yd_i + (1 - \alpha) Xd_i \tag{37}$$

where Xd_i is the Mahalanobis distance for sample i to predict, α is a weighting factor and Yd is a normalized chemical distance calculated as follows:

$$Yd_i = \frac{\left|y - y_{i,l}\right|}{\max\left(\left|y - y_{i,l}\right|\right)} \tag{38}$$

where y, the estimated unknown property estimated by a global regression and $y_{i,l}$ is the reference value of each local sample.

The next step consists of developing the local model and predicting the new sample. CARNAC is particular because it does not use any regression method, but the prediction of the new sample is determined by a weighted average of the y values of the selected samples. The logarithm of the similarity index is used as a weighting factor. LOCAL and LWR methods, however, predict the y value of the new spectrum using a regression method (PLS or PCR). In LOCAL, several models are developed for different PCs and the final prediction is a weighted average of all predictions. LOCAL algorithm automates the choice of the number of PCs. The weight associated with each prediction is:

$$W_k = \frac{1}{X_{residual} \times RMS_{BetaCoef}} \tag{39}$$

where W_k is the weight associated with the kth PC, $X_{residual}$, the difference between the reconstructed spectra (from PLS or PCR decomposition) and the true spectrum to be predicted, and $RMS_{BetaCoef}$, the root mean square of the regression coefficient. The best principal component will present the best reconstructed spectra and the smoothest regression coefficient and will thus maximize the weighting factor ($X_{residual}$ decreases while $RMS_{BetaCoef}$ increases with the number of PCs). For LWR, the determination of the best PC is not automated. However, a weighting function is applied to each selected sample using a cubic weight function (figure 16) based on the distance of the sample to predict with the local calibration samples (the farthest is scaled to have a distance of 1).

The weighting function is as follows:

$$W_S = \left(1 - d^3\right)^3 \tag{40}$$

where W_S is the weight associated with a calibration sample and d, the scaled distance between the new sample to predict and the calibration sample.

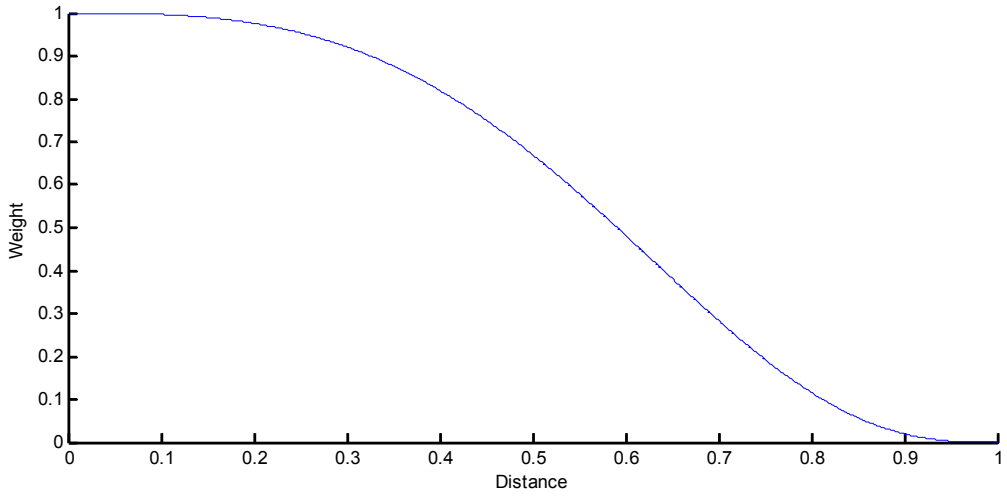


Figure 16: Cubic weight function.

Like every calibration procedure, the last step in the development of a local method should be a validation procedure. All validation strategies presented in section 2.3.5. are applicable to local models.

Local strategies have their pros and cons. They can be as accurate as global models, robust, and able to deal with non-linear relationships. However, they are time consuming, have little instrumental support (mostly offline applications), and cannot be approved in heavy regulatory environments. From a local model point of view, no calibration can be approved because each model is different. However, this regulatory issue could be solved by approving the sample database, since the database is fixed for each new sample (Cogdill et al, 2002).

2.2.5. Standardization

Calibration models are often developed using spectra of a single instrument called a master unit. This master unit should be chosen among other units of the network of instrument available to the chemometrician because of its precision and accuracy (Siska et al, 2001). However, the master unit is often located in the research laboratory and does not participate to the actual online processes or quality controls. The model needs to be transferred to these instruments. This is called standardization and is an essential part of the implementation of NIRS.

Three types of standardization techniques exist:

- Optical standardization or calibration transfer by adaptation of the secondary unit's spectra to match master's spectra
- Post regression correction or standardization by correction of the predicted values of the secondary units using calibration developed on master unit.
- Robust models or standardization by calibration model adaptation

2.2.5.1. Optical standardization techniques

Optical standardization techniques try to adjust secondary unit's spectra to the spectra of the master unit. The intent is for the calibration model to perform as well as if a sample was scanned on the master unit. To perform the spectral modifications, a set of samples, called standardization samples, need to be run on both master and secondary units. Bouveresse and Massart (1996) presented a comprehensive description of the selection and the use of standardization samples in their review of standardization techniques. This can be done with a high leverage method proposed by Wang et al (1991) or the Kennard-Stone algorithm (Kennard et al, 1969). Bouveresse et al (1994) presented two simple techniques to select standardization samples. The first consisted in selecting samples that were the best predicted among the validation set. The second consisted in using samples

from a different source but of similar nature (not artificial samples). When dealing with several prediction models, using the same standardization samples for all factors is often convenient, but may not result in the best standardization since a sample may be predicted well for a factor but not for another.

Several techniques exist to match spectra. Five of the best known techniques will be presented here. Other methods have been developed, but remain marginal in their use because their lengthy calculations do not bring real improvements compared to existing methods. Examples of complex techniques are ANN based standardization (Despagne et al, 1998; Duponchel et al, 1999), maximum likelihood PCA (Andrews et al, 1997), positive matrix factorization (Xie et al, 1999), Kalman and Wiener filters (Teppola et al, 1999 and Siska et al, 2001 respectively), and other standardization of the regression coefficient techniques (Wang et al, 1991).

2.2.5.1.1. Single wavelength standardization

This method, developed by Shenk and Westerhaus (Paynter et al, 1983; Shenk et al, 1985), consists of correcting, for each wavelength, the absorbance shifts (adjust the intensity differences between instruments). It is done by regressing, one at a time, the absorption values of a standardization set scanned on the master unit against the one obtained when the same set is scanned on the secondary unit. A slope and an offset are obtained and the secondary unit is corrected. This simplistic technique is applicable only when a limited shift in the **X**-axis (wavelength) exists.

Figure 17 shows the application of this technique in standardizing two Foss Infratec units (master unit: Infratec 1241, secondary unit: Infratec 1229). Twenty soybean standardization samples were used and the wavelength of interest was 852 nm. More information about the material can be found in Chapter 3.

The secondary unit had a lower intensity in its measurements and the correction function at this wavelength would be:

$$W_{Scor}^{852} = 0.88 \times W_{S}^{852} + 0.19 \tag{41}$$

where W_{Scor}^{852} corresponds to the corrected value of the secondary unit at 852 nm, W_{S}^{852} is the original value of the secondary unit, and 0.88 and 0.19 are the slope and the offset values respectively obtained when fitting a first order polynomial to the data.

The quality of the standardization samples is critical because a wrong approximation of the correction factors at each wavelength can have a large impact on the final SEP.

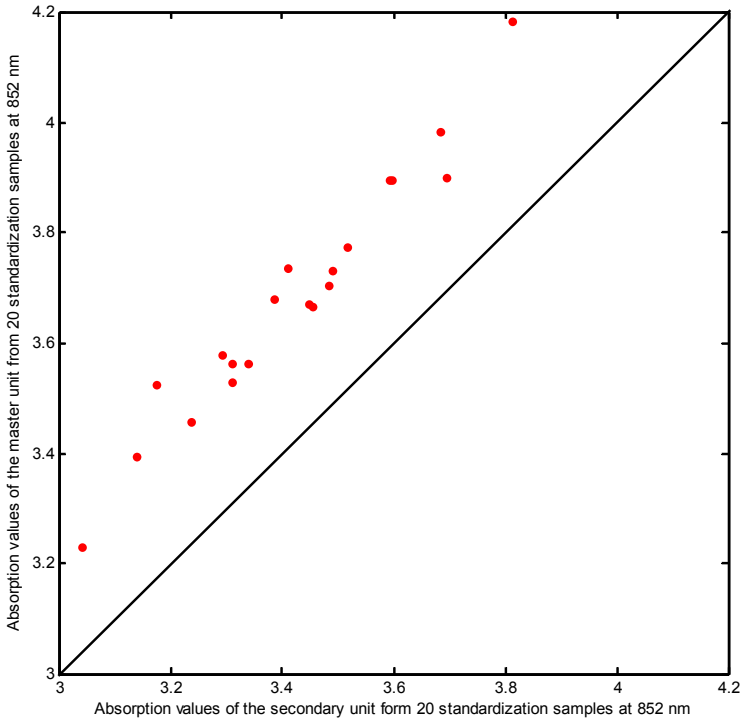


Figure 17: Single wavelength standardization. An offset and a slope need to be applied to this wavelength to correct the signal.

2.2.5.1.2. The patented algorithm – Shenk and Westerhaus

An updated version of the single wavelength standardization included the possibility to correct for shifts in the **X**-axis. This method has been patented and is used by the chemometric package WinISI (US Patent 4866644 – September 1989). Shenk et al (1993) and Bouveresse et al (1994) describe the functioning of the algorithm.

It is divided into two steps. The first step is called wavelength index correction and consists in correcting the **X**-axis shifts while the second part corrects the intensity differences. Standardization spectra are preprocessed using a first derivative treatment. For each wavelength of the master, a spectral window of neighboring wavelength on the secondary instrument is chosen and for each wavelength of the window, the correlations with the master are computed. A quadratic model is fit to the correlation values to estimate more precisely the position of the wavelength that produces maximum correlation. The fit is between the wavelength that has the highest correlation and its two neighboring wavelengths. The new locations obtained from the quadratic model (inflection points of the windowed quadratic equation) are recorded and a new spectrum is built. A second quadratic model is developed to relate the master wavelength to the matching wavelength on the modified secondary unit spectra. Definitive values for the secondary unit wavelengths corresponding to the master wavelengths are obtained. This process is called the wavelength index correction. The modified secondary unit spectra are then interpolated. Wavelengths of the secondary unit are shifted to the corresponding master wavelength, at each wavelength (similar to what was implemented in the single wavelength standardization method) using a linear regression. A slope and an offset for each wavelength are obtained and are used to correct spectral intensity.

The wavelength index and the spectral intensity correction factors are stored and applied to new spectra scanned on the slave instrument. This method is adapted to many standardization situations involving similar instruments. When facing instruments with more complex differences like a peak broadening, the patented algorithm is not sufficient

because it "assume that no relationship exists between neighboring correction models" (Feudale et al, 2002).

2.2.5.1.3. *Direct standardization*

Wang et al (1991) introduced two standardization techniques able to cope with peak broadening. Both methods imply that spectra from the master and the secondary unit are linearly related and this relation can be described by a transformation matrix such as

$$\mathbf{S_M} = \mathbf{S_S F} \tag{42}$$

where $\mathbf{S_M}$ is a spectra scanned on the master unit, $\mathbf{S_S}$ a spectra scanned on the secondary unit, and $\mathbf{F}$ the transformation matrix.

In direct standardization (DS), the transformation matrix is simply estimated as

$$\mathbf{F} = \mathbf{S_S}^{+}\mathbf{S_M} \tag{43}$$

where $\mathbf{S_S}^{+}$ is the pseudo inverse of $\mathbf{S_S}$. With $\mathbf{F}$ calculated, any new spectra $\mathbf{S_N}$ can be modified to the original measurement space so that the calibration will predict it appropriately using

$$\mathbf{S}_{Ntrans} = \mathbf{S}_N \mathbf{F} \tag{44}$$

where $\mathbf{S_{Ntrans}}$ is the new spectrum modified to the original measurement space. $\mathbf{S_{Ntrans}}$ is then applied to the calibration model developed on the master unit.

The computation of $\mathbf{F}$ assumes that the difference between instruments is due to instrumental variations. However, the variation in the chemical composition (due to non-complete sample homogeneity) is also modeled and may be a source of error. Also, because the number of samples used to create $\mathbf{F}$ (standardization samples) is smaller than the number of channels to evaluate, DS is subject to overfitting. $\mathbf{F}$ is typically estimated using PCR or PLS to obtain a least squares solution.

Another approach to limit overfitting is to reduce the number of channels estimated at a time. This is the purpose of piecewise direct standardization (PDS).

2.2.5.1.4. Piecewise direct standardization

In DS, each wavelength of the master unit is related to all the wavelengths of the secondary unit. PDS is a local alternative to DS. Influence of shifts of the **X**-axis between instruments is limited to certain spectral regions and does not impact the standardization of other wavelengths.

PDS performs the same calculations as DS but at a local level. The user defines a window size and transformation coefficients are calculated to relate one wavelength of the master unit with several wavelengths on the secondary unit:

$$\mathbf{r}_j = \mathbf{R}_j \mathbf{b}_j \tag{45}$$

where $\mathbf{r}_j$ is the absorption value at wavelength *j* from the master unit, $\mathbf{R}_j$ is the absorption value at wavelength *j* from the secondary unit, and $\mathbf{b}_j$ is the vector of transformation coefficients for wavelength *j*. The size of the window usually varies between 3 and 5 wavelengths and a wavelength on the secondary unit is used several times to explain different wavelength on the master unit (moving window, one master wavelength at a time). For each window, transformation coefficients are estimated and assembled to form a banded diagonal matrix **F** according to

$$\mathbf{F} = diag\left(\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}}, ..., \mathbf{b}_j^{\mathrm{T}}, ..., \mathbf{b}_k^{\mathrm{T}}\right) \tag{46}$$

where k is the number of wavelengths.

Thus, the transfer matrix relates the response of a number of wavelengths (size of the window) of the secondary unit to a single wavelength of the master unit (figure 18).

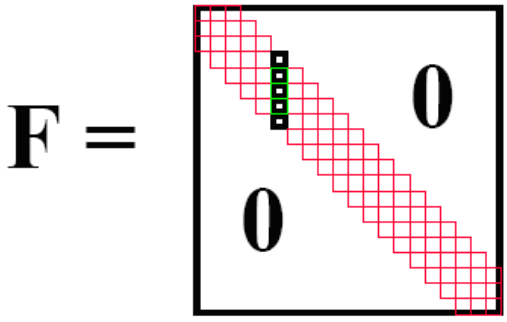


Figure 18: Structure of the transformation matrix (Wise et al, 2006).

PDS is often used as reference method for other standardization techniques. Many authors have used it successfully, but few instruments are able to use a transfer matrix onboard (Perten DA (Perten Instruments AB, Huddinge, Sweden) can use PDS with a window size of 1). The standardization using DS or PDS has to be done offline. This situation is similar to the use of the patented algorithm except that Foss NIRSystems instruments are equipped with software that can accommodate the standardization parameters. Notice that a PDS with a window size of 1 is equivalent to single wavelength standardization. The difference is that in PDS, coefficients are estimated using a PCR or PLS.

To carry on the example introduced with single wavelength standardization, table 5 presents the transformation matrix for wavelength at 850 nm, 852 nm, and 854 nm when a window size of 3 is used on the secondary unit. With each wavelength a bias is associated. In this particular case, they are -0.0114, -0.0074, and -0.0077 for 850 nm, 852 nm, and 854 nm respectively. With these parameters, the calculation of the corrected absorption value of a spectrum collected on a secondary unit for the wavelength at 852 nm would be:

$$W_{Scor}^{852} = 0.3601 \times W_{Sraw}^{850} + 0.3601 \times W_{Sraw}^{852} + 0.3589 \times W_{Sraw}^{854} + (-0.0074) \quad (47)$$

The reason for coefficients to be so similar is that there is no shift in the wavelength axis.

Table 5: Structure of the transformation matrix for the example using a window size of 3 wavelengths.

		Master Unit			
		850	**852**	**854**	**…**
Secondary Unit	**850**	0.5420	0.3601		
	852	0.5401	0.3601	0.3614	
	854		0.3589	0.3601	…
	856			0.3589	…
	…				…

An attempt to use DS and PDS on spectra transformed in the wavelet domain was published by Tan et al (2001). They showed that the standardization at the approximation level with PDS and the detail level with DS was more robust and reliable than using DS or PDS only.

2.2.5.2. Post regression correction techniques

This second type of model transfer method does not modify the spectra. It adjusts predictions made on the secondary unit by the model developed on the master unit. Post regression correction techniques are easily implemented by every embedded software or firmware and are also easier to understand (for users).

Often called slope and bias correction, this technique is widely used among very similar instruments (one can note the misuse of the term bias which in reality corresponds to term intercept). In this technique, a standardization set is scanned on all instruments (the secondary units as well as the master units) and predictions are compared to the reference values (Bouveresse et al, 1996). A linear regression is implemented and the slope and intercept of the trend line are determined. These parameters are then used to systematically correct predictions from the secondary instrument using the prediction equation developed on a master unit. This is a very simple approach that requires only that the range of the standardization samples cover the range of the calibration samples.

A special case of slope and intercept correction is bias only correction. Bias, as defined by equation 31, is calculated from standardization samples predictions and reference values and is added to the future predictions. This situation is suitable only when the slope is significantly equal to one. This signifies that the difference between the two units is mainly based on absorption intensities differences and not a shift in the **X**-axis. The slope in post regression correction appears when the wavelength alignment between the instruments is not correct.

As an example, predictions of the standardization set by a protein model developed on Infratec 1241 and run on Infratec 1229 are compared with the reference measurements. Figure 19 shows the relationship between the two units as well as the slope and intercept necessary to implement slope and bias correction and the bias for bias only correction.

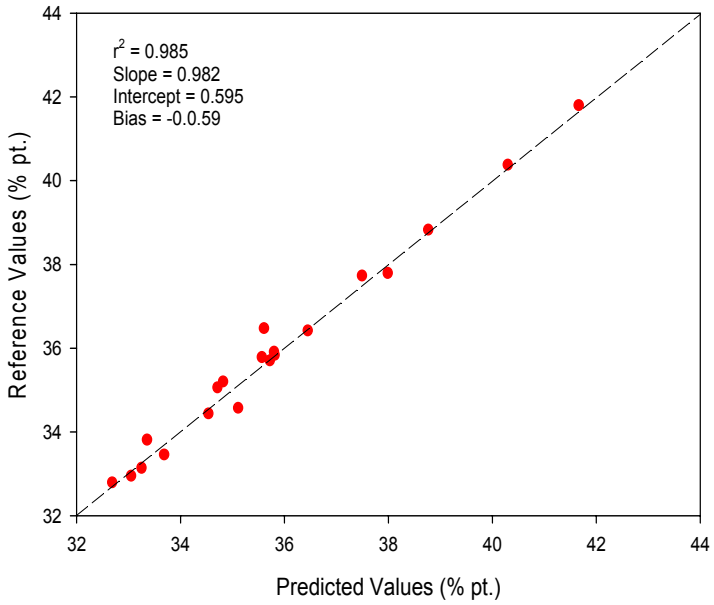


Figure 19: Post-regression correction results for protein. Slope and intercept or bias only correction can be used to correct the measurements.

In this situation, using slope and bias post regression correction, the prediction on the secondary unit should be corrected as follows:

$$\hat{X}_{Cor}^{Sec} = 0.982 \times \hat{X}_{Raw}^{Sec} + 0.595 \quad (48)$$

And using bias only correction, it would look like:

$$\hat{X}_{Cor}^{Sec} = \hat{X}_{Raw}^{Sec} + (-0.059) \quad (49)$$

2.2.5.3. Robust standardization techniques

Robust statistics are statistics that do not rely as much as classical statistics on assumptions. Many of the classical techniques require a normal or uniform distribution of the data. It is the case of MLR, PCR, and PLS while ANN and SVM are robust methods. They are more tolerant to outliers and emulate classical methods. However, the term robust, in robust standardization techniques, represents the idea of being resistant to outliers, but does not always involve the use of robust statistic techniques.

Several robust standardization methods have been developed, but few papers have been published on the topic.

2.2.5.3.1. Spectral preprocessing

A possibility to make a calibration robust is to use pretreatment methods to remove instrumental interferences. But these methods do not work if there is a shift in the **X**-axis among instruments (Perten DA instruments uses MSC to standardize diode array instruments). Standardization by preprocessing is fairly easy, but cannot cope with large or inconsistent instrumental differences.

2.2.5.3.2. *Variable selection*

It is also possible to make a model robust by selecting wavelengths that are not subject to shifts (isonumeric wavelengths (Mark and Workman, 1988)) and to develop models using MLR. However, it is necessary to correct for intensity differences (Dean and Isaksson, 1993).

2.2.5.3.3. *Including external variability*

This family of robust methods consists in scanning all or part of the calibration set on all the instruments of the network and by developing a calibration based on all the spectra. This way, the calibration algorithm does the standardization work by adapting the model to each instrument. Fearn (2001) noted that "The resulting calibration will almost certainly be less accurate than a calibration for a single instrument, but this may be a price worth paying". Often preprocessing methods are used to help remove the instrumental interferences.

2.2.5.3.4. *Orthogonal-based techniques*

These methods are based on the statistical theory that the column space of the **X**-matrix (set of all possible linear combination of column vectors) is "the sum of two subspaces, among which only one contains information useful for the model "(Roger et al, 2003). Thus, by doing the appropriate projection, one can develop a model based only on the adequate **X**-matrix. There exists two ways of estimating the "parasitic" subspace. The first consists of finding the space orthogonal to y. This is what orthogonal signal projection (OSC) and orthogonal projections to latent structures (O-PLS) are doing. The second approach estimates the space in which the external factor occurs. Transfer by orthogonal projection (TOP), dynamic orthogonal projection (DOP), and error removal by orthogonal subtraction (EROS) are the most popular techniques to find and remove the orthogonal space.

- Orthogonal signal correction

OSC was introduced by Wold et al in 1998. The idea is to compute loading weights, **w**, such that the scores **t** calculated from **t** = **Xw** describe as much variance as possible. The decomposition into scores is similar to the one used to develop PLS. The constraint is that this variance is not correlated to y. It is possible to repeat the process and remove more orthogonal factors from the **t** matrix. This corrected **t** matrix is then used to develop a PLS model.

- Orthogonal projections to latent structures

Trygg and Wold introduced a variation of OSC in 2002. Instead of removing the orthogonal signal from the **t** matrix prior to calibration development, the O-PLS method removes the orthogonal information from the PCs calculated by PLS. The reconstructed **X**-matrix can then be used to develop a calibration model. O-PLS was reported to perform better than OSC.

OSC and O-PLS are, as mentioned, in section 2.3.3, preprocessing methods. Their primary goal is to remove noise. Authors reported that systematic variations could be successfully removed by the methods. These variations can be due to scattering or baseline effects, but from a standardization standpoint, the removal of systematic variations can correspond to deleting instrumental signatures and thus improve model transferability.

- Transfer by orthogonal projection

Andrew and Fearn (2004) published an orthogonal technique aiming to remove the interfering components from the **X**-matrix in the situation of calibration transfer. Standardization samples are measured on both master and secondary units. A difference spectrum is obtained by subtracting standardization samples from both units and a PCA is performed on the difference matrix **D**. The first *k* loadings of **D** are used to form the matrix

P representing the direction of main variation between units. **P** is orthogonalized on the **X**-matrix to obtain a corrected **X**-matrix ($\mathbf{X_{corr}}$) for between-instrument differences using

$$X_{corr} = X\left(I - PP^T\right) \tag{50}$$

where **I** is the identity matrix. $\mathbf{X_{corr}}$ is then used to develop calibration models.

TOP requires the use of standardization samples. But in certain situations, collection of samples is not possible and orthogonalization is performed using the DOP method.

- Dynamic orthogonal projection

In DOP (Zeaiter et al, 2006), standardization samples are replaced by virtual standards. These standards are chosen to represent a variation to be removed. In calibration transfer situations, a few samples collected from the secondary instruments (X_t) are used to create virtual standards using a kernel function. Equation 51 demonstrates how virtual standards are created using the calibration set (**X** and y) and the reference value of the samples collected on the secondary unit (y_t):

$$\hat{X}_t = AX \text{ with } a_{ij} = F_{y_{ti}}\left(y_j\right) \tag{51}$$

where $\hat{X}_t$ are the virtual standards, $F_{y_{ti}}$ is a Gaussian kernel function centered on y_{ti} for the i^{th} sample and the j^{th} variable. A difference matrix **D** is calculated ($D = \hat{X}_t - X_t$) and similarly to TOP, the first *k* loadings from a PCA are selected to form **P**. The **X**-matrix is then orthogonalized similarly to TOP using equation 50.

By not requiring the same samples run on both master and secondary unit, DOP is more flexible, especially for on-line applications.

- Error removal by orthogonal subtraction

EROS (Zhu et al, 2008) is based on the same principles as TOP and DOP. The difference arises from the way the **P** matrix is calculated. In EROS, the difference matrix **D** is mean centered and a new matrix **W** representing the difference between measurements is calculated as follows:

$$W = \sum_{i=1}^{m} \frac{D_i D_i^T}{(r-m)} \tag{52}$$

where **r** is the total number of spectra in the **D** matrixes and *m* is the number of samples (also the number of **D** matrixes). **W** represents the pooled within-sample covariance matrix of the replicate spectra. The *k* first PCA loadings of **W** are selected and used to form the matrix **P**, used to orthogonalize the **X**-matrix using equation 50.

Orthogonal methods need to be tuned. The number of times orthogonal components are removed for OSC and O-PLS need to be optimized. Similarly, TOP, DOP, and EROS require tuning *k*, the number of factors to orthogonalize.

TOP, DOP, and EROS have the advantage of being embedded in the calibration model (the regression algorithm will not take into account the orthogonal signal and the beta coefficients will emphasis the orthogonalization by not taking into account the orthogonal part of the signal in prediction mode). Thus, these methods are implemented during the calibration process. Validation samples do not need any orthogonalization step on the contrary to all other preprocessing and standardization methods. They have the tremendous advantage to be useable by all instrument software settings (even firmware). These same methods can be used to remove other interferences than instrumental differences such as temperature, batch effects, and other unwanted external effects presenting a repetitive effect across samples (Zeaiter et al, 2006).

2.2.5.4. Summary

Choosing amongst the different standardization methods presented here is often guided by the capacities of the instrumentation. While post-regression correction and some orthogonal methods can be used by all instruments, very few can actually implement DS, PDS, and advanced preprocessing methods. When working off-line, all methods are available, but the competency of the user is challenged when not guided by any software suite. Only the patented algorithm does not require any input from the user (except spectra). The software performs all calculations and provides an "easy to use" standardization file. Other optical, post regression, and robust methods rely on the capacity of the programmer to write the appropriate code (in Excel (Microsoft Corporation, Redmond, WA, USA), in MATLAB (The MathWorks, Inc., Natick, MA, USA), or in R (The R Foundation for Statistical Computing, Vienna, Austria)) and to tune the parameters. Even post regression methods can be tedious to use, because using a slope different from one may have important consequences (under-predicting low values and over-predicting high values, or the contrary). It also acknowledges the fact that a possible shift in the wavelength axis exists between both units.

In addition, it may be necessary to combine standardization methods (PDS to correct shift of the wavelength and bias only correction to increase accuracy). Finally, the quality of the prediction model is of importance. A calibration model presenting limited accuracy and precision will be difficult to transfer to secondary units without increasing the error.

Anthony M.C. Davis, a pioneer of NIRS wrote, in 2007, in a forum dedicated to diffuse reflectance spectroscopy: "If it ain't broke. It don't need fixing" which was here to remind the users that implementing a calibration transfer implies that the results it provides are better than they were before (when applying the model directly to the secondary unit spectra). The use of standardization methods is not systematic; however, a control of the calibration models is necessary.

2.2.6. Conclusions

In this second chapter, we introduced the challenges that traceability constitutes, in terms of risk management and instrumentation. Near infrared spectroscopy was presented as a tool of choice for traceability because of its quick, non-destructive, cheap, and manipulation-free analysis. Basic physical and chemical theories were discussed as well as typical instrument settings. The treatment of the spectral data and the issue of instrument standardization were presented.

Fearn (2001) noted the limited literature available on the comparison of standardization methods and never discussed the standardization among instrument brands. At the time this literature review was created, only three studies had been published on the subject (Heckman et al, 1987; Shenk et al, 1993, and Wang et al, 1993). The need for being able to transfer historical databases to newer technologies (i.e. diode array detection, LED light source) is tremendous. Because of the complexity of the differences between technologies, many techniques presented here may have difficulties to integrate all the instrumental differences. Many instrumentation companies have developed methods to use databases from other instrument brands on their machines (Perten DA 7200 can use spectra from NIRSystem units, FOSS North America, Eden Prairie, MN, USA), but very few have been published.

In the following chapters, we will investigate the use of commonly used standardization methods in a situation of inter-brand calibration transfer (two brands). This will help to fill the lack of literature in the comparison among standardization methods. Then, robust methods will be tested using spectral preprocessing methods in the spatial and the Fourier and wavelet domains. These approaches will be compared with the development of new local chemometrics techniques (selection of both samples and variables) and the implementation and the modification of orthogonal methods, in an effort to increase robustness of calibration models (removing from the X-matrix the information bringing instability) when performing in both intra and inter-brand situations.

CHAPTER 3.

MATERIAL AND METHODS

This third chapter describes databases, validation strategies, and algorithms used in this dissertation. In a first section, samples, instruments, and reference analysis are presented. In a second section, the experimental design is outlined as well as the various methods tested from the literature, developed during the doctoral research, and the underlying framework structuring this dissertation. Finally, software used to perform calculations is described.

3.1. Spectral data

3.1.1. Samples

Approximately 630 whole soybean samples scanned during crop years 2002 to 2005 were included in the calibration set. These samples were collected over the years by the Grain Quality Laboratory of Dr. Charles R. Hurburgh Jr. at Iowa State University, Ames, Iowa. They represent the genetic, climatic, and environmental variability encountered in the USA. They have been collected over the years through national surveys, variety trials, and service operations.

3.1.2. Instruments

All samples were scanned on four near infrared spectrometers. The network of instruments was composed of two Foss Infratec Grain Analyzers (FOSS North America, Eden Prairie, MN, USA): Infratec 1229 (s/n: 553075) and Infratec 1241 (s/n: 12410350) and two Bruins OmegAnalyzerGs (Bruins Instruments, Puchheim, Germany): OmegAnalyzerG (s/n: 106110 and 106118).

These instruments were chosen because both models (Infratec, OmegAnalyzerG) are approved by the National Type Evaluation Program (NTEP) and are entitled to perform legal analysis in the USA. They are transmittance units with a spectral range from 850 nm to 1048 nm with an increment of 2 nm. OmegAnalyzerGs scan 741 wavelengths from 730 nm to 1100 nm, but a software option allows the reduction of the spectra to the same 100 datapoints as the Infratec units. This is the instrumental set up approved by NTEP.

A fixed pathlength of 30 mm was used. Samples were run at room temperature. Each sample was analyzed simultaneously on the four instruments. The number of samples was not always the same for all four units; a sample identified as an outlier for a unit could be well predicted by another unit. Also, not all instruments had the possibility to scan the smaller samples without changing the number of subsamples collected (Infratecs 1229 collected 10 subsamples, Infratec 1241 collected 10 subsamples, OmegAnalyzerGs collected 16 subsamples). With small samples, OmegAnalyzerGs and Infratec 1241 stop the analysis and ask for refill while Infratec 1229 reports an error. This situation was considered closer to a field experiment than reducing to exactly the same samples in all four calibration sets.

A subsample corresponds to a single reading of the absorption of a sample. A spectrum is generally composed by the average of several subsamples. The number of subsamples is set by the manufacturers or the Grain Quality Laboratory for being optimal in terms of spectral quality and time needed to perform the analysis.

3.1.3. Validation sets

In order to evaluate calibration model performances in terms of precision and accuracy, two validation sets were created. A first set of 20 samples chosen from the calibration set was used to determine the ability of the various models to predict the sample variability present in the calibration set. The 20 samples were not included in the calibration set used to develop models. This set was called “set of known variability”.

A second set of 40 samples from crop year 2006 was chosen to determine the ability of the various models to predict new variability. These samples underwent different environmental conditions than those in the calibration set and the genetic present may be different. The 40 samples were not included in the calibration set used to develop models. This set was called “set of unknown variability”.

3.1.4. Reference methods

The present study considered three soybean parameters. Protein content was determined by combustion (AOAC 990.03); oil content was determined by ether extract (AOCS Ac 3-41); both by Eurofins Scientifics, Inc., Des Moines, IA, USA. Linolenic acid content was determined by gas chromatography using the method described by Hammond (1991), in the Department of Agronomy at Iowa State University, Ames, Iowa. Summary statistics for the calibration and validation sets are presented in table 6.

Table 6: Summary statistics of the calibration and validation sets.

Parameter	Instruments (serial)	n	Average Concentration (%)	Range (%)	Standard Deviation (%)
Protein (13% moisture basis)					
	Infratec 1229 (s/n: 553075)	638	36.06	29.74 - 46.50	2.80
	Infratec 1241 (s/n: 12410350)	624	36.04	29.74 - 46.50	2.76
	OmegAnalyzerG (s/n: 106110)	628	36.10	27.48 - 46.50	2.86
	OmegAnalyzerG (s/n: 106118)	625	36.09	27.48 - 46.50	2.92
	Validation set 1	20	36.72	30.65 - 45.43	3.32
	Validation set 2	40	37.61	33.78 - 41.92	2.14
Oil (13% moisture basis)					
	Infratec 1229 (s/n: 553075)	629	18.42	11.89 - 22.52	1.78
	Infratec 1241 (s/n: 12410350)	635	18.40	11.89 - 22.52	1.80
	OmegAnalyzerG (s/n: 106110)	655	18.51	11.89 - 22.52	1.73
	OmegAnalyzerG s/n: (106118)	638	18.38	11.89 - 22.48	1.80
	Validation set 1	20	18.55	15.27 - 21.68	1.57
	Validation set 2	40	18.46	13.82 - 21.63	1.83
Linolenic Acid (% of oil)					
	Infratec 1229 (s/n: 553075)	741	6.21	0.89 - 11.08	2.53
	Infratec 1241 (s/n: 12410350)	713	6.30	0.90 - 11.08	2.48
	OmegAnalyzerG (s/n: 106110)	701	6.19	0.89 - 11.08	2.56
	OmegAnalyzerG (s/n: 106118)	696	6.20	0.89 - 11.08	2.56
	Validation set 1	20	4.31	1.03 - 10.56	3.52
	Validation set 2	40	6.12	1.72 - 8.94	1.91

3.2. Experimental design and methodology

The present dissertation aims to test existing methods and develop new approaches to solve the calibration transfer issue for near infrared spectrometers. Using the four instruments of the selected network, four calibration transfer strategies were implemented.

In a first step, instrumental differences between the four units of the network will be studied and a master unit for each network will be chosen. Then, a comparison of common standardization methods will be presented. These results will be compared with the implementation of (i) new spectral preprocessing methods based on the modification of high frequency components in the Fourier and wavelet domains, (ii) new variable selection and local chemometrics techniques, and (iii) orthogonal methods.

3.2.1. Benchmark models – each instrument on its calibration set

All four instruments of the network were first calibrated on their own calibration set. This allowed the selection of a master unit for each brand and also the determination of benchmark results for comparison with other techniques since these results represented the best performances obtainable from each instrument, using standard calibration practices. These original models, as well as every other model developed in this study, used partial least squares regression. Spatial pretreatments were also common to all calibrations (second derivative (Savitzky-Golay – 25-point window and 5^{th} order polynomial for protein, 3^{rd} order polynomial for oil, 5-point window and 3^{rd} order polynomial for linolenic acid), normalization to unit area under the curve, and autoscaling (zero mean, unit standard deviation)). Prediction performances were evaluated in terms of precision with SEP and RPD and accuracy with bias. RPD was often preferred to SEP because of its resolution power. For a small change in SEP, a larger change in RPD can be observed and allow a better comparison between methods.

Master units were then used for the development of intra-brand and inter-brand standardization strategies. In intra-brand situations, models of the master units were

transferred on the secondary unit of the same network while in inter-brand situations, master unit models were transferred onto the two units of the other instrument brand.

3.2.2. Common calibration transfer techniques

A first approach compared six common calibration transfer strategies from the literature.

These methods were:

1. Signal filtering in the spatial domain – Robust method
2. Piecewise Direct Standardization (PDS) – Optical method
3. Direct Standardization (DS) – Optical method
4. Joint master calibration sets – Robust method
5. Slope and Bias – Post-regression correction
6. Bias only – Post-regression correction

Standardization samples used by four of these methods were taken from the calibration set (before creation of the models). Ten subsets of twenty samples each were randomly selected. Their protein, oil, and linolenic concentration ranges were checked to match the ranges of the remaining calibration samples. New random subsets were created if the ranges were not adequate. A calibration for each parameter was developed without one of the subset at a time and the subset taken apart was predicted with the calibration (similar to block cross validation). The subset that presented the lowest SEP and bias for all three parameters (equally weighted) was chosen as standardization set. It was then removed from the final calibration set.

The ability of these six methods to transfer the model of master units to the secondary unit of their network and the two units of the other network brand was evaluated. While for methods 1, 2, and 3, only the predictions of secondary units were modified, methods 4, 5, and 6 modified the predictions of the master units as well as the

secondary units. Joint calibration set regrouped the calibration set of both master units and, as a consequence, the prediction performances of the master units were modified. Similarly, post-regression corrections being done against reference measurements on the standardization samples, predictions of the master units were also adjusted.

Method 1 – Signal filtering in the spatial domain – featured the use and the combination of spectral preprocessing methods for the reduction of spectral differences between units and brands. All four categories of spectral pretreatment methods were used: baseline correction (smoothing, derivative, detrending, and weighted least squares baseline), normalization and scattering correction (normalization, standard normal variate, and multiplicative scatter correction), interference removal (orthogonal signal correction and generalized least squares weighting), and scaling methods (mean-centering and autoscaling). All meaningful and reasonable combinations were evaluated (n = 75). Only the five best combinations were reported.

Finally, the use of PDS involved the choice of a window size on the secondary unit spectra for which absorption values of the master unit were approximated. Window sizes of 1, 3, 5, 7, and 9 were tested and only the results with the best settings were reported.

3.2.3. Frequency components filtering

This second approach featured three new ways to filter near infrared spectra. They are spectral pretreatment methods used prior to spatial pretreatment methods.

3.2.3.1. Fourier smoothing

The Savitzky-Golay algorithm was used to develop a smoothing filter for the high frequency components on outputs from a fast Fourier transform. The smoother was applied to a range of Fourier index representing high frequency components. An iteratively search of the best range of Fourier coefficients that could be altered by the smoothing process and

still provide similar of better performances than on non filtered data was performed. Frequency filtered spectra were obtained after an inverse fast FT and were used to develop calibration models. The possibility to apply the smoother to the validation sets or not was investigated.

A filter with a window size of 25 points and a 5th order polynomial was applied to the periodigram. Savitzky-Golay smoothing was preferred to moving average and boxcar methods because they tend to preserve spectral features (Savitzky and Golay, 1964).

Fourier smoothing can be summarized as follows:

1. Transform data into the Fourier domain
2. Apply smoother to high frequency components
3. Inverse fast Fourier transform

with the optimization process consisting of finding the appropriate filter parameters (window size, polynomial order) and the starting point (first Fourier index at which the smoother is applied). Note that since all spectra consisted of 100 absorption values, the last Fourier coefficient corresponded to the 50th sine coefficient.

3.2.3.2. Fourier fitting

A cubic polynomial was used to fit the magnitude components of the high frequency components from a FT of the calibration set. Fourier coefficients were converted into magnitude and phase angle. The third order polynomial was fit to the magnitude components of a range of Fourier coefficients representing high frequency components. Similarly, as in Fourier smoothing, the search of the starting point is an iterative process. The last point of the fit was always the 50th sine coefficient. The possibility to apply the fit to the validation sets or not was also investigated.

The polynomial coefficients were found by fitting a cubic polynomial on the average spectrum of the twenty standardization samples. Polynomial fitting was chosen to remove local features and contrast with the Savitzky-Golay smoothing.

Fourier fitting can be summarized as follows:

1. Determine a fit pattern (mean spectrum of standardization set)
2. Transform fit pattern and calibration sets to the Fourier domain
3. Create polynomial on magnitude signal from fit pattern
4. Apply fit to calibration set
5. Inverse fast Fourier transform

with the optimization process consisting in finding the appropriate fit pattern, polynomial order, and starting point.

3.2.3.3. Wavelet smoothing

The Savitzky-Golay algorithm was used to develop a smoothing filter for the high frequency components on outputs from a wavelet transform using a Daubechies 4 wavelet. The detail components of the first level decomposition were smoothed. The smoother used a 3-point window and a 2nd order polynomial. The size of the windows was dictated by the shape of the signal; while FT high frequency components were very smooth, WT detail components were very noisy. Frequency filtered spectra were obtained after an inverse WT and were used to develop the calibration models.

Wavelet smoothing can be summarized as follows:

1. Calculate detail components from first level decomposition
2. Apply smoother to the detail components
3. Inverse wavelet transform

with the optimization process consisting in finding the appropriate filter parameters (window size, polynomial order).

Fourier smoothing, fit, and wavelet smoothing preprocessed spectra were then preprocessed in the spatial domain with second derivative, normalization, and autoscaling (similar setting as benchmark calibration models). The ability of these three methods to transfer the model of the master unit to the secondary unit of their network, and the two units of the other network brand was evaluated. It was possible to update and compare

benchmark results by applying these high frequency-based smoothing methods in the situation of developing and validating each instrument on its own calibration samples.

3.2.4. Robustness enhancement by sample and variable selection

This third approach to calibration transfer featured the use of local chemometrics methods as well as variable selection and their combinations.

3.2.4.1. Sample selection

Along with LWR, LWRY, and LOCAL, two new approaches, with a similarity index based on Euclidian distance among Fourier coefficients at a specific Fourier index, were developed.

3.2.4.1.1. Fourier distance index for local similarity

A new method was implemented to select similar samples to the samples to predict for the development of local regression models. Sample selection is based on the determination of an index of similarity that orders calibration samples in relation to the new sample to predict. In LOCAL, sample similarity is based on the coefficient of correlation between spectra. In LWR, the mahalanobis distance between principal component scores is measured. A new index of similarity was calculated by measuring the Euclidian distance between the phase angles of the spectrum to predict and the calibration set spectra. Samples are first transformed into the Fourier domain with a fast Fourier transform algorithm and distances were measured at a specific Fourier index. The selection of the Fourier index is an iterative process. In this study, the fifths cosine coefficient was chosen because it belongs to the first pair of parameters to not undergo low frequency variations from sample to sample.

The number of samples to include in the local model, as well as the number of principal components to include in the PLS model is chosen by the end user. Selected

samples were weighted using a cubic weight function similar to the one used in the LWR algorithm.

3.2.4.1.2. Fourier distance index for local similarity semi automated

Fourier Distance Index for Local Similarity Semi Automated (FDILS_SA) is an alternative to FDILS where the choice of the number of principal component to include is automatically determined. In the same fashion as LOCAL, a weight function is applied to the predictions of a range of selected PCs and the median of the weighted predictions is chosen for final output.

The weight function is based on the behavior of the root mean square error of calibration (RMSEC) and the RMSECV that should respectively go lower and lower and go higher and higher, with an increasing number of PCs. The difference between the first derivative of RMSEC and the RMSECV was calculated and used in the weight function

$$\hat{y}_W = median\left(\hat{y}_i \times \left(1 + W_i\right)\right) \tag{53}$$

where $\hat{y}_W$, the weighted prediction is a function of $\hat{y}_i$, the predicted value for PC *i* and the difference of first derivatives W_i at the same PC.

FDILS and FDILS_SA can be summarized as follows:

1. Transform sample to predict and calibration set in the Fourier domain
2. Measure Euclidian distance between sample to predict and calibration set samples based on the 10th Fourier coefficient (mean value included)
3. Sort sample of the calibration set by increasing distance
4. Select only the kth samples following user inputs
5. Apply cubic weight function and develop PLS model

FDILS	FDLIS_SA
6. User choose the appropriate PC	6. Calculate weight function
7. Output predicted value	7. Output median value of weighted predictions

The ability of these four local chemometrics methods to transfer the model of master units to the secondary unit of their network and the two units of the other network brand was evaluated. It was also possible to update and compare benchmark results by applying these local chemometrics methods in the situation of developing and validating instruments on their own calibration samples.

3.2.4.2. Variable selection

The ability of GA and PSO to select spectral features was evaluated. In addition, a modification to PSO was proposed where p_{id} (the best state of the particle so far) was replaced by the state of the best neighboring particle. The use of neighboring particles is a protection against overfitting. This modification, suggested by Kennedy and Eberhart (2001), was named Neighboring PSO (NPSO). Lines 15, 16, and 17 of the pseudo code were then modified into

$$v_{id}(t) = W_t \times v_{id}(t-1) + \varphi_1\left(n_{id} - x_{id}(t-1)\right) + \varphi_2\left(g_{id} - x_{id}(t-1)\right) \tag{54}$$

where n_{ip} represents the state of the best neighboring particle of particle i for variable d.

For all three methods, the optimization process was performed iteratively by modifying the number of initial individuals/particles, the number of selected variables in each, and the crossover/mutation rates and the learning factors for GA and PSO respectively. It was possible to update and compare benchmark results by applying these techniques in the situation of developing and validating instruments on their own calibration samples.

3.2.4.3. Combination of feature and sample selection

For completeness, sample selection and variable selection were combined. Variable selection was used prior to sample selection to mimic the use of filter instruments. Again, it was possible to update and compare benchmark results for all units.

3.2.4.4. Wavelength matching

To cope with possible differences in the wavelength axis among units (in inter-brand situations), three wavelength matching methods were implemented. The first method was a piecewise direct standardization with windows size of one which consisted in scanning standardization samples on all units and correcting each secondary unit absorption values with a slope and an offset based on a linear regression of both master and secondary units absorption for the specific wavelength. The second method consisted in a simplistic copy of the patented algorithm of Shenk and Westerhaus (1985), where each variable of a transfer set from the master unit was matched up with the variable in the neighborhood of the same variable on the secondary unit (window size of 5) that had the closest absorption value as illustrated by figure 20. The variable in black on the master unit was approximated by the closest gray variable (crossed) from available variables from the secondary unit. On this specific example, a wavelength shift of 2 nm can be observed. The third method implemented post regression correction where validation set predictions on the secondary units were corrected with either a slope and offset or a bias calculated by doing a regression on the standardization set predicted values of the same model from both the master and the secondary units.

The ability of these three spectral matching methods to improve calibration transfer was evaluated. To limit calculations, only the best spectral matching method was implemented in combining feature and sample selection.

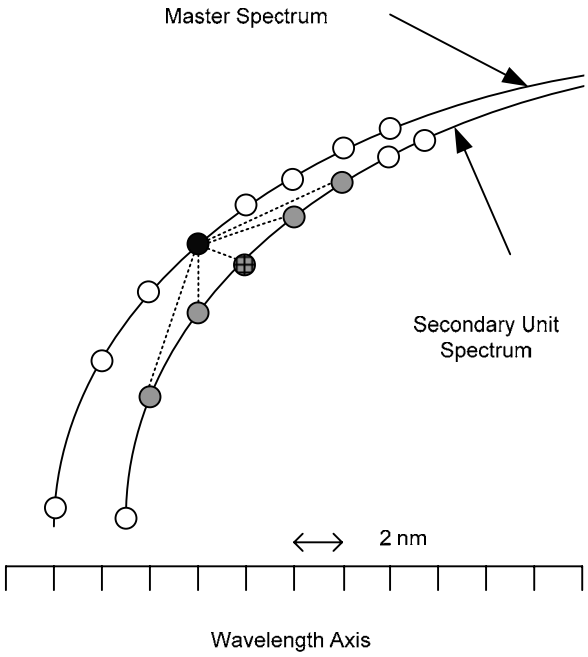


Figure 20: Variable matching for wavelength axis selection by closest selection.

3.2.5. Orthogonal based robustness enhancement methods

This fourth, and last, approach to the enhancement of calibration robustness employed orthogonal methods. OSC, O-PLS, TOP, DOP, and EROS were compared in their ability to remove the influence of instrumental differences. Standardization samples were employed to form the matrix **P** used to orthogonalize the **X**-matrix prior calibration for TOP, DOP, and EROS.

A modification to DOP was also evaluated. Instead of taking into consideration only the parameter of interest (protein or oil) when creating the virtual standards, both parameters were used either at the same time or one after the other to form the matrix ***D***. They were respectively named DOP 2y block and DOP 2y sequential. In DOP 2y block, the ***D*** matrix was formed by concatenating difference spectra for protein and for oil while in DOP 2y sequential, the calibration set was first orthogonalized for differences in protein, then for differences in oil. This was an attempt to increase the robustness of models by considering more variability than present for one parameter alone.

This process was done in to ways:

- Orthogonalization of **X** for one parameters then for the other (method named DOP 2y sequential)
- Orthogonalization of **X** for both parameters at the same time (method named DOP 2y block)

Since samples in the calibration set were collected over four years, an attempt to model instrumental changes was done. Instrument aging differently, if the inter-annual variability could be removed, the transfer of calibration models might be made easier. A set of twenty samples scanned every year by the Grain Quality Laboratory for control purposes, different from calibration, validation, and standardization sets, were used to perform TOP and EROS. These samples had the particularity to have been scanned the same years calibration samples were collected and reflected instrumental variations over the years. They were used to form the difference matrix ***D***. They were called TOP 4years and EROS 4years.

3.2.6. Conclusions

The comparison between common standardization methods and the various attempts of robustness enhancement methods developed in this study (spectral filtering in Fourier and walevet domain, variable or/and sample selection, and orthogonal based methods) will provide the NIRS community with a consequent library of results for grain based calibration model transfer.

To help the reader, table 7 summarizes all calculations performed in this dissertation. For each approach, instruments for which standardization methods can be implemented are displayed. Also, benchmark updates are notified when each instrument can be calibrated on its own calibration set. To simplify the table, notations 1229, 1241, 106110, and 106118 were used for Infratec 1229, 1241, and OmegAnlyzerG 106110 and 106118 respectively. Also, Infratec 1241 and OmegAnalyzerG 106118 comport a tag (star) on the first row meaning they were designated network masters (see chapter 4. for more details).

Table7: Framework of the dissertation. Calibration and standardization performances are calculated when instrument's serial numbers are displayed.

		Foss Infratec Network						OmegAnalyzerG Network					
Original		1229		1241*				106110		106118*			
Litterature comparizon	Method	Spatial pretreatment only	Piecewise Direct Standardization	Direct Standardization	Join Calibration set	Slope & Bias	Bias only	Spatial pretreatment only	Piecewise Direct Standardization	Direct Standardization	Join Calibration set	Slope & Bias	Bias only
	On Own Calibration Set	1229 1241 106110 106118			1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118			1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118
	Model Transfers	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110	1229 1241 106110	1229 1241 106110	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118
Signal Filtering	Method	Fourier Smoothing		Fourier Fit		Wavelet Smoothing		Fourier Smoothing		Fourier Fit		Wavelet Smoothing	
	On Own Calibration Set	1229 1241 106110 106118		1229 1241 106110 106118		1229 1241 106110 106118		1229 1241 106110 106118		1229 1241 106110 106118		1229 1241 106110 106118	
	Model Transfers	1229 106110 106118		1229 106110 106118		1229 106110 106118		1229 1241 106110		1229 1241 106110		1229 1241 106110	
Sample and Variable Local Modeling		Sample Selection					Variable Selection			Sample Selection					Variable Selection		
	Method	Fourier Distance Index for Local Similarities	Fourier Distance Index for Local Similarities_ Semi Automated	Locally Weighted Algorithm	Locally Weighted Algorithm Y	LOCAL	Genetic Algorithms	Particle Swarm Optimization	Neighboring Particle Swarm Optimization	Fourier Distance Index for Local Similarities	Fourier Distance Index for Local Similarities_ Semi Automated	Locally Weighted Algorithm	Locally Weighted Algorithm Y	LOCAL	Genetic Algorithms	Particle Swarm Optimization	Neighboring Particle Swarm Optimization
	On Own Calibration Set	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118
	Model Transfers	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118
	Spectral Matching	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118
Orthogonalization	Method	Orthogonal Signal Correction	Orthogonal Projections to Latent Structures	Transfer by Orthogonal Projection x2	Dynamic Orthogonal Projection	Dynamic Orthogonal Projection 2Ys x2	Error Removal by Orthogonal Subtraction x2	Orthogonal Signal Correction	Orthogonal Projections to Latent Structures	Transfer by Orthogonal Projection x2	Dynamic Orthogonal Projection	Dynamic Orthogonal Projection 2Ys x2	Error Removal by Orthogonal Subtraction x2
	Model Transfers	1229 1241 106110 106118	1229 1241 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 1241 106110 106118	1229 1241 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118	1229 106110 106118

3.3. Software

All calculations were performed with MATLAB R2008a (The MathWorks, Natick, MA, USA). The PLS_toolbox v. 4.2.1 was used to develop all calibration models, GA, and spatial preprocessing (Eigenvector Research, Wenatchee, WA). Coefficients of the cubic polynomial in the Fourier fit filtering method were determined with the Curve Fitting Toolbox v. 1.2. Statistical analyses were performed in JMP v. 6.0.0 (SAS Institute Inc., Cary, NC).

CHAPTER 4.

RESULTS AND DISCUSSION

In this chapter, results of calibration development and transfer will be presented in the order in which they were introduced in the previous chapter: common standardization methods, filtering in Fourier and wavelet domains, variable and local chemometrics optimization, and orthogonal based methods. But before, instrumental differences are detailed and network masters selected.

4.1. Instrumental differences across the network on NIR units

Figure 21 shows the difference between the average spectra of the standardization samples of instrument of the same brand and of both brands. Spectra were pretreated with a second derivative treatment (Savitzky-Golay 5-point window, 3rd order polynomial) to enhance differences and remove baseline effects.

Within the Infratec network (figure 21.a.), we could observe a difference in absorption intensities at almost every peak. No shift in the wavelength axis was visible, but some peaks appeared to have a different area (peak **6** or **8**). OmegAnalyzerG spectra (figure 21.b.) were very similar with only slight differences from intensity absorption. In figure 21.c., Infratec 1241 and OmegAnalyzerG 106118 were compared. A wavelength axis shift, a difference in absorption intensity, and differences in peak areas could be observed. In terms of variation in the **X**-axis, there was no shift at peak **1** and **2**, a shift of 2-wavelengths at peaks **3**, **4**, **5**, **6** and a 4-wavelength shift at peaks **7** and **8**. In addition to this non-linearity, differences in peak areas could be observed at peaks **3** and **5**.

Figure 22.a. shows a PCA score plot of the calibration sets of Infratec 1241 and OmegAnalyzerG 106118 preprocessed with second derivative (25 point windows, 3rd order

polynomial) only. The two instruments could be easily differentiated with samples from Infratec 1241 in the right cluster. Instrumental differences could be successfully reduced when additional preprocessing methods were applied (second derivative - 25 point windows, 3rd order polynomial, normalization to unit area and scaled to mean zero, unit variance); these clusters were reduced (figure 22.b.).

Information present in one instrument could thus be used to predict unknown samples from the second instrument. This further supports the possibility to implement inter-brand standardization scenarios using spectral preprocessing (spatial and Fourier, wavelet domains) and also local methods.

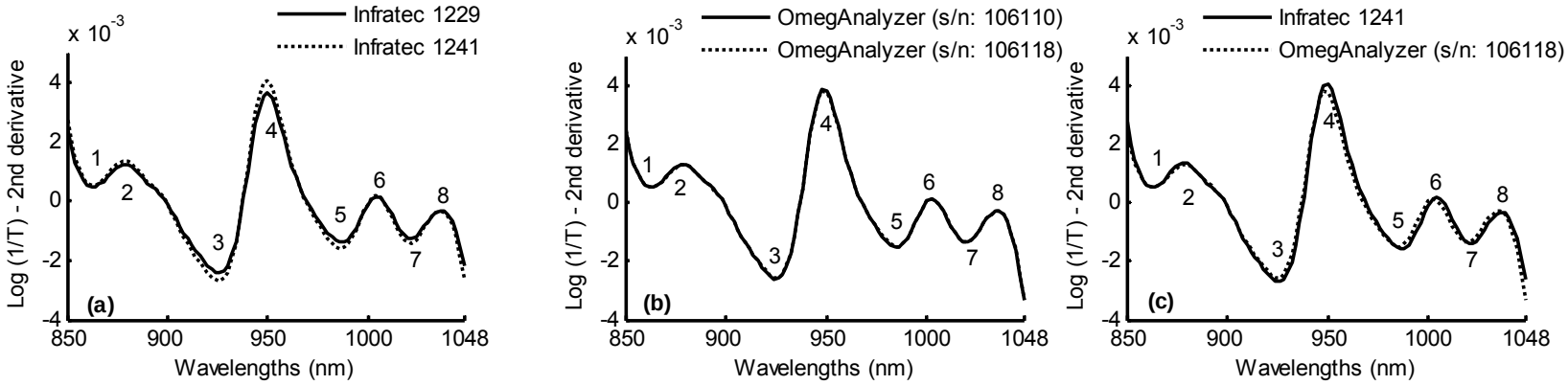


Figure 21: Spectral differences among and across instrument brands. Raw scans of standardization samples were averaged and preprocessed with second derivative (Savitzky-Golay 5-point window, 3rd order polynomial).

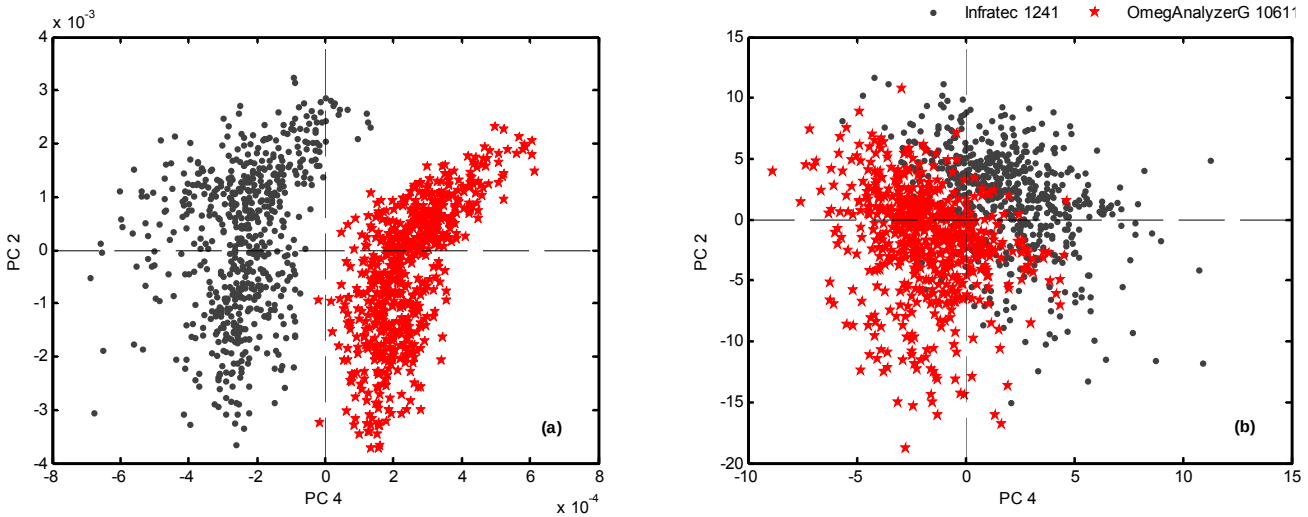


Figure 22: Principal component score plots of the calibration sets of Infratec 1241 (crosses) and OmegAnalyzerG 106118 (squares). (a) represents PCA scores on unpreprocessed spectra. (b) represents PCA scores on preprocessed spectra (second derivative, normalization)

4.2. Master unit selection for each instrument brand

Table 8 presents validation results when each instrument was calibrated on its own calibration set. To select master units, a method proposed by Fearn (1996) and originally created by Pitman (1939) to compare regression techniques when predicting the same validation set was used. This method compares the two sets of prediction errors with a t-test to determine the significance of the difference.

For the prediction of the first validation set (of known variability), Foss Infratec 1241 and Bruins OmegAnalyzerG (s/n: 106118) were significantly more precise within their brand ($\alpha = 0.05$). For the second validation set, results were similar across instruments except for the prediction of oil where OmegAnalyzerG (s/n: 106110) predicted significantly lower contents than OmegAnalyzerG 106118 ($\alpha = 0.05$). In terms of accuracy, results were more heterogeneous. It is interesting to notice that while for protein and oil, no clear pattern could be observed within each instrument brand, Foss Infratec 1241 and Bruins OmegAnalyzerG (s/n: 106118) gave significantly lower biases for linolenic acid.

Because of their better precision and equivalent accuracies, Foss Infratec 1241 and Bruins OmegAnalyzerG (s/n: 106118) were designated masters of their respective instrument brand network.

Table 8: Validation results of calibration models developed on their own calibration set.

Instruments (Serial)	Parameters	Validation set 1 (n = 20) (known variability)			Validation set 2 (n = 40) (unknown variability)		
		Protein 13%mb	Oil 13%mb	Linolenic acid %Oil	Protein 13%mb	Oil 13%mb	Linolenic acid %Oil
Infratec 1229	RPD	10.41	4.98	4.02	3.89	4.22	0.88
(s/n: 553075)	Bias (% pt.)	0.16	-0.62	0.20	0.11	-0.44	-1.45
Infratec 1241	RPD	11.25	6.52	4.31	4.64	4.97	1.18
(s/n: 12410350)	Bias (% pt.)	0.20	-0.25	0.33	0.06	-0.40	-1.52
OmegAnalyzerG	RPD	9.29	5.17	3.77	4.17	5.55	0.94
(s/n: 106110)	Bias (% pt.)	0.12	-0.16	1.31	0.17	-0.48	-1.53
OmegAnalyzerG	RPD	11.41	5.90	4.26	4.54	5.35	0.98
(s/n: 106118)	Bias (% pt.)	0.12	-0.23	0.66	0.16	-0.47	-1.38

4.3. Evaluation of common standardization methods

4.3.1. Intra-brand calibration transfer

Validation results from the six standardization methods were compared to when each instrument was calibrated on its own calibration set. Infratec 1241 and OmegAnalyzerG (s/n: 106118) were masters of their respective networks.

Five spectral pretreatment combinations gave significantly higher RPDs than all other combinations. Since they were not significantly different, their results were averaged. They were:

- Second derivative (25 point window, 3rd order polynomial) + normalization (unit area) + autoscaling
- SNV + Second derivative (25 point window, 3rd order polynomial) + normalization (unit area) + autoscaling
- MSC + Second derivative (25 point window, 3rd order polynomial) + normalization (unit area) + autoscaling
- Second derivative (25 point window, 3rd order polynomial) + normalization (unit area) + OSC + autoscaling
- Second derivative (25 point window, 3rd order polynomial) + GLSW + normalization (unit area) + autoscaling

Notice that the first combination was already the set of pretreatment methods that was used to develop benchmark calibrations. This reinforces the validity of the initial calibration process choices.

Figure 23 shows, for each validation set and each product, RPDs obtained from the various standardization methods.

Differences in precision existed between the two validation sets. Except for oil, validation results for protein and linolenic acid were significantly lower when predicting the second set than with the first. Validation results for linolenic acid on validation set 2 were poor. The limited range and precision of the models when applied to a new variability probably explain these validation statistics. However, this could signify that an element which is hard to calibrate might be harder to standardize consistently due to the poor performance of the calibration on standardization samples.

For the validation set 1, DS and PDS gave significantly lower RPDs than other methods ($\alpha = 0.05$). The other four standardization methods and the original individual models were not significantly different. Preprocessing methods used in the model development (similar for original models and models) resolved the baseline absorption differences between units of the same brand. All standardization methods (except optical techniques) gave equal or better results than when the instrument was calibrated on its own database. Thus, it is possible to improve the prediction ability of an instrument by transferring from a higher quality model. Choosing a network master is one of the most important steps in the constitution of a network.

For the validation set 2, DS and PDS were significantly poorer than other methods for the prediction of protein but not for oil and linolenic acid. There was no difference among other standardization methods, and the original models. Except for oil, predictions of transferred models were again equal or better than those from models developed on an instrument's own database.

In all cases, the simplest methods (joint calibration sets, post regression correction and spatial preprocessing) gave equivalent results to the original instrument models. No simple standardization method was significantly better than the others.

Surprisingly, DS and PDS gave the lowest RPDs. The implementation of these methods requires that the exact same sample be scanned on both instruments. The variability within the samples could be solved as suggested by Bouveresse et al. (1994) by

using sealed cup to ensure that the scans of the standardization samples correspond on each instrument to exactly the same sample. However, from instrument to instrument, the use of such cups is not always possible (which was our case). The increase of the error for DS/PDS could also be due to the estimation of additional parameters (the **F** matrix) which inevitably adds error. This is consistent to the methods used to estimate **F** since DS, which involves only singular value decomposition, performed globally better than PDS, which requires the development of numerous local PCR or PLS models.

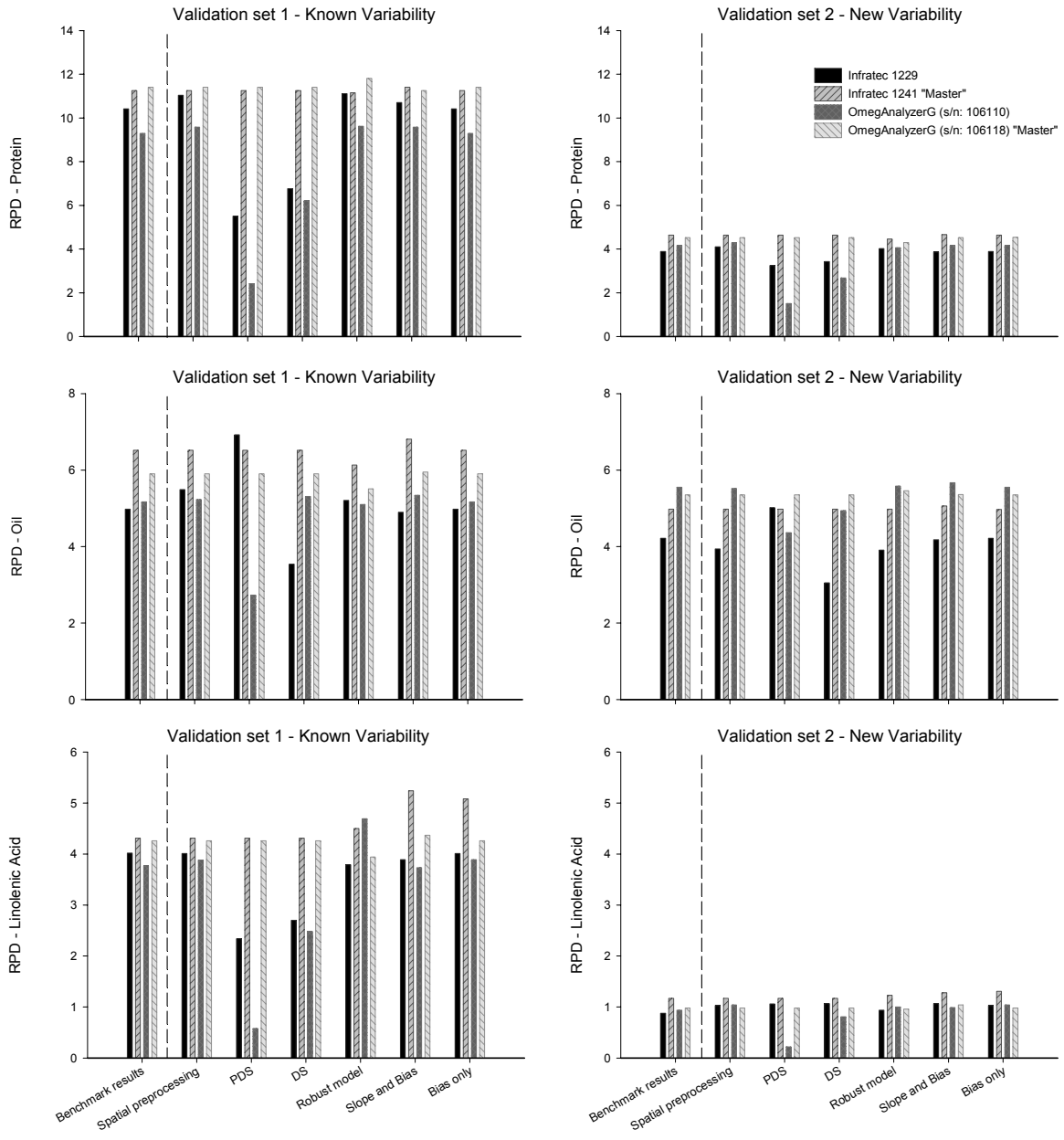


Figure 23: Validation of the different standardization methods in an intra-brand situation.

4.3.2. Inter-brand standardization

Figures 24 and 25 present the prediction results when the calibration models of the two masters were transferred to the two instruments of the other brands (Infratec 1241 to OmegAnalyzerG instruments and OmegAnalyzerG (s/n: 106118) to Infratec instruments respectively).

Again, PDS and DS gave significantly lower RPDs while there were no significant differences between the other four standardization methods ($\alpha = 0.05$). For protein and oil, among the four methods that performed well (spectral preprocessed models, robust models, slope and bias, and bias only), three – robust models, slope and bias and bias only – gave equal or better results than when each instrument was calibrated on its own calibration set.

For linolenic acid, the results were different because for validation set 1, only robust models gave satisfactory results. In both situations (Infratec and OmegAnalyzerG masters of the network), PDS, DS, and slope and bias methods gave significantly poorer results than other techniques. When Infratec 1241 was master of the network, "spectral pretreatment" and "bias only" methods were not significantly different from the original results. When OmegAnalyzerG (s/n: 106118) was master, only the robust model was statistically similar to original results. For validation set 2, when Infratec 1241 was master, PDS and slope and bias gave significantly lower RPDs than other methods (consistent with what was observed for set 1). However, when OmegAnalyzerG (s/n: 106118) was master of the network, there were no significant differences among standardization methods.

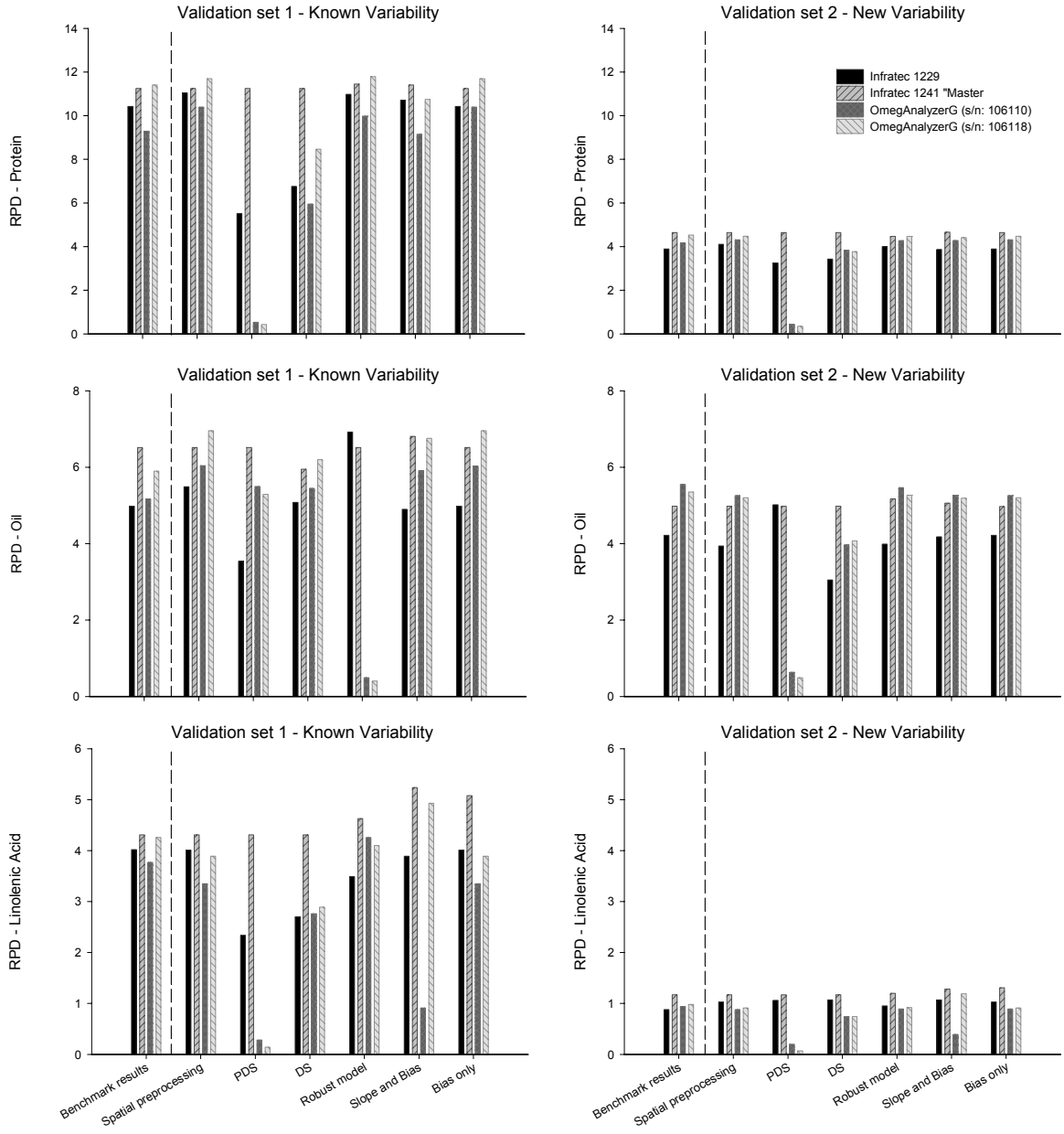


Figure 24: Validation of the different standardization methods in an inter-brand situation with Foss Infratec 1241 master of the network.

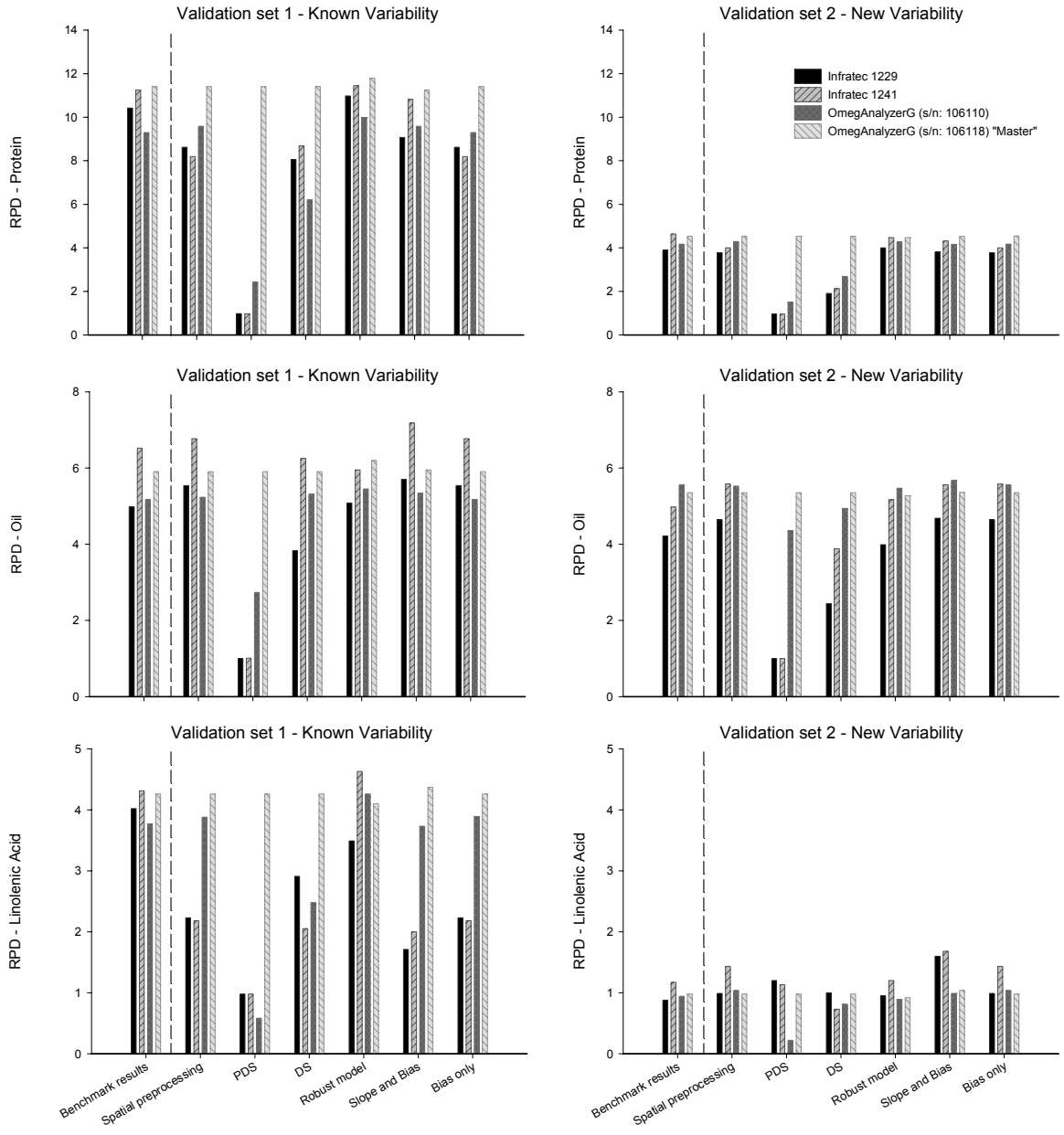


Figure 25: Validation of the different standardization methods in an inter-brand situation with Bruins OmegAnalyzerG (s/n: 106118) master of the network.

4.3.3. Standardization methods precision and accuracy

The standardization process should preserve a precision similar or better than when a model is developed on each instrument. AACC method 39.00 allows in increase of 0.5 x SEP between the master and the secondary unit, which implies a 10% decrease in precision. In the present study, no statistical difference existed among standardization methods except for the optical techniques.

Calibration transfer should also output a bias as close as possible to the observed bias on the master unit. Table 9 presents the average bias per standardization method and per parameter, among master and secondary units, for intra-brand situations. Biases for inter-brand were similar to those of intra-brand. There were no significant differences between biases of masters and secondary units. Standardization methods had an impact on the precision but not the bias, in the conditions of this study ($\alpha = 0.05$). This would imply that two transfer methods, one for precision and one for accuracy may be appropriate.

Table 9: Average biases for intra-brand standardization situation.

	Validation set 1			Validation set 2		
	Protein (% pt)	Oil (% pt)	Linolenic acid (% pt)	Protein (% pt)	Oil (% pt)	Linolenic acid (% pt)
Master units						
Benchmark results	0.16	-0.24	0.48	0.11	-0.43	-1.45
Robust models	0.11	-0.21	0.29	0.08	-0.44	-1.47
Slope and offset	0.23	-0.21	0.40	0.18	-0.41	-1.47
Bias only	0.09	0.27	0.57	0.04	-0.46	-1.30
Secondary units						
Benchmark results	0.14	-0.15	0.75	0.14	-0.45	-1.49
Spectral preprocessing	0.24	-0.32	-0.08	0.25	-0.55	-2.09
Robust models	0.11	-0.21	0.29	0.08	-0.44	-1.47
PDS	-0.13	-0.08	-1.17	0.18	-0.56	-1.52
DS	0.20	-0.16	1.70	0.03	-0.44	-1.51
Slope and offset	0.23	0.10	0.77	0.25	-0.40	-1.26
Bias only	0.07	-0.20	-0.93	0.07	-0.51	-2.94

4.3.4. Conclusions

The transfer of calibration from Foss Infratec to Bruins Instruments OmegAnalyzerG was possible. Except for optical methods, standardization methods were able to cope with the various instrumental differences between brands. The simplest methods performed well in both intra and inter-brand situations. Thus, post-regression correction methods and robust models provided the most satisfactory results, of the most commonly used methods. In terms of accuracy, transferred models gave similar biases than those calculated for the master units. Post-regression correction methods, while providing results as precise as original validation results, were, for the most part, not able to erase the bias. The example of linolenic acid, a parameter difficult to measure by both near infrared spectroscopy and the reference method, showed that a model with poor performances is also difficult to transfer.

These results show that consistency can be achieved by the appropriate methods. Even though the two brands in this study were of similar hardware design, differences in absorption intensities and wavelength axis shifts were overcome.

4.4. Frequency components filtering

The previous section showed the difficulty of applying standardization methods to unstable prediction models such as linolenic acid calibrations. From this point on, this parameter was given up since it would not bring anything new from the calibration transfer point of view. The study carried on with protein and oil.

4.4.1. Smoothing and fitting method performances in calibration

Figure 26 compares validation results for precision when each instrument was calibrated on its own calibration set (validation set not filtered). "Original" represent results obtained with spatial preprocessing only while "Fourier Smoothing", "Wavelet Smoothing", and "Fourier Fitting" results represent the use of the various frequency modification methods in addition to spatial preprocessing methods.

No significant difference existed between original results and the various Fourier and wavelet domain filtering-based methods. Among these methods, wavelet performed significantly worse than the two others for oil and validation set 2. No difference could be observed between filtered or not filtered validation samples. However, an improvement of calibration precision was observable in 81% of the situations (except for OmegAnalyzerGs oil – validation set 1 and Infratec 1241 protein – validation set 2).

The modification of frequency components in addition to spectral preprocessing appears to be insensitive to year to year variability. The information removed was most likely to be noise and not signal. These methods appeared highly dependent on the parameter and the instrument.

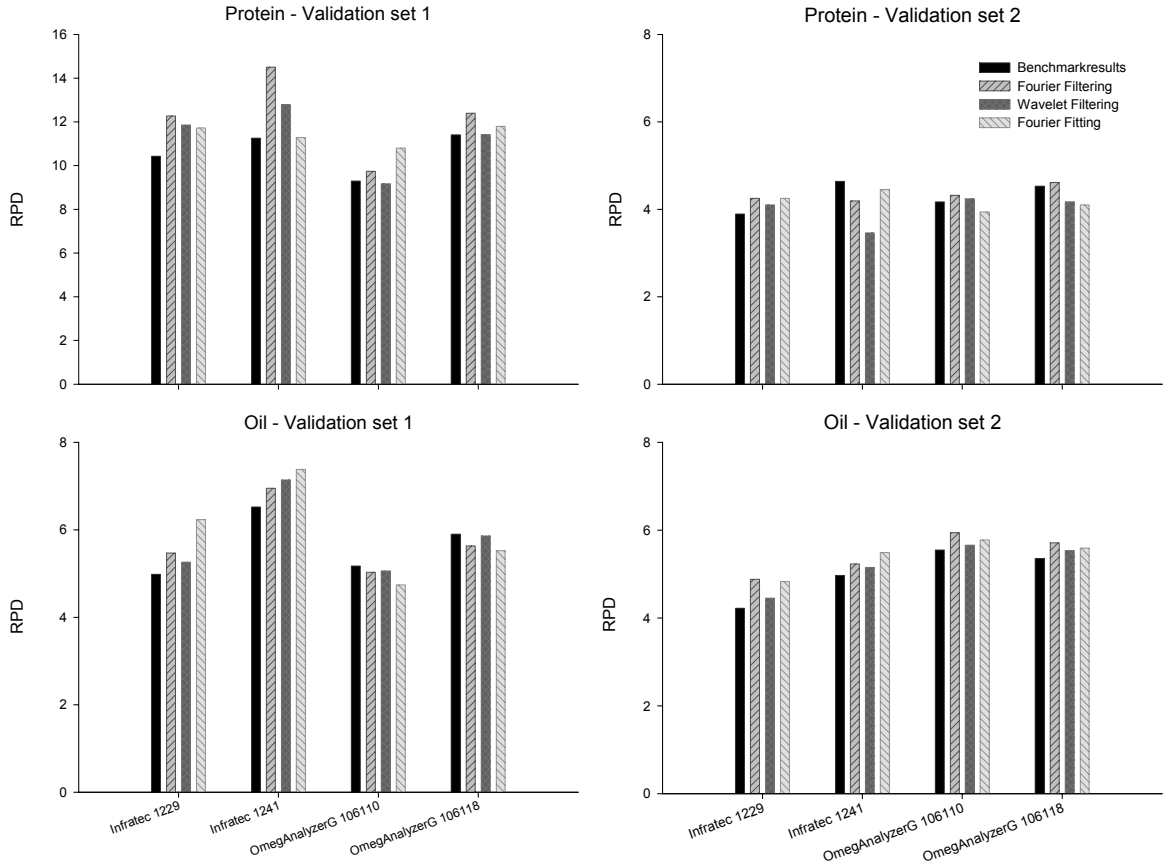


Figure 26: Effect of smoothing and fitting methods in calibration situation.

4.4.2. Intra-brand standardization situation

Predictions models of each instrument's masters were transferred to the secondary unit of the respective brand. Master model predictions were compared with original predictions and with models transferred with and without frequency modifications. Figure 27 presents the results (when validation sets were not filtered).

All secondary unit predictions were improved by one or more transfer methods but not significantly. In several situations, transferred models performed better on the secondary unit than on the master unit. This may signify that the noise source is similar

from instrument to instrument of the same brand. Master and secondary unit predictions were significantly different for validation set 2 (protein and oil).

These results show that the appropriate processing of the spectra can, in most cases, avoid implementing additional standardization methods. However, results tend to show that it is not possible to generalize the use of preprocessing methods (both spatial and frequency based). Their use remains an iterative process, instrument and parameter dependent. Also, differences between master and secondary units were not completely removed even when secondary unit precisions were improved.

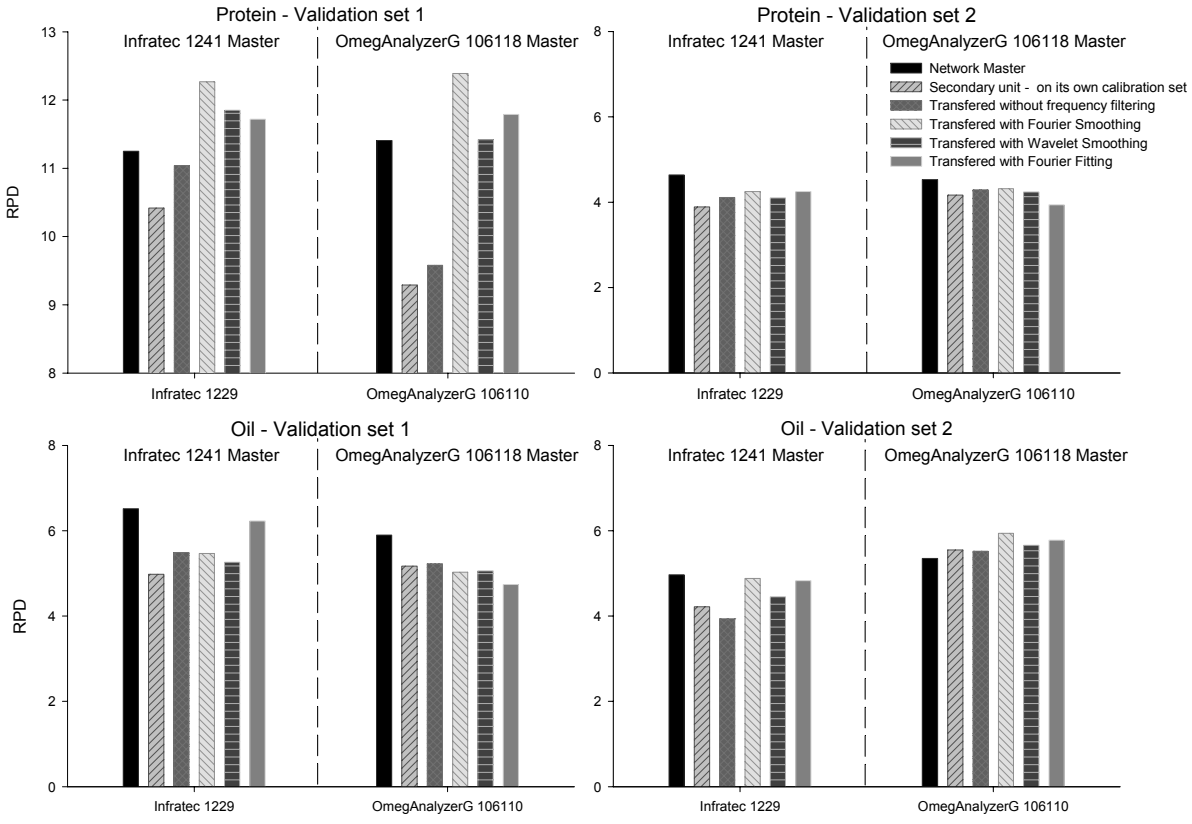


Figure 27: Evolution of secondary unit precision in intra-brand situation. Infratec 1241 was master unit of Infratec 1229. OmegAnalyzerG 106118 was master unit of OmegAnalyzerG 106110.

4.4.3. Inter-brand standardization

Models of the master of each instrument brand were transferred onto the two instruments of the other brands. Figure 28 compares each master's performances with secondary unit performances when models were transferred with and without frequency pretreatment (validation sets were not filtered).

With Foss Infratec 1241 master of both Bruins units, Fourier filtering and fitting gave significantly higher RPDs than wavelet filtering for protein. There was no difference between master and secondary units. The process of applying all three filtering methods to the validation set did not provide better results. An improvement of the secondary unit performances was observable in all situations relative to models developed on their respective calibration sets. The use of frequency filtering and fitting methods was suitable to transfer calibrations from Infratecs to Bruins. In some situations, transferred models performed better than the master unit.

With Bruins OmegAnalyzerG 106118 master of the inter-brand network, similar results were obtained than when Infratec 1241 was master. Wavelet filtering performed poorly for protein, and filtering validation samples did not make a difference. However, the difference between the OmegAnalyzerG master unit and the secondary units was not removed for protein and validation set 2. The use of frequency filtering and fitting methods was again suitable to transfer calibrations from OmegAnalyzerGs to Infratecs with secondary units performing better than the network master. However, the use of these methods remained parameter and instrument dependent.

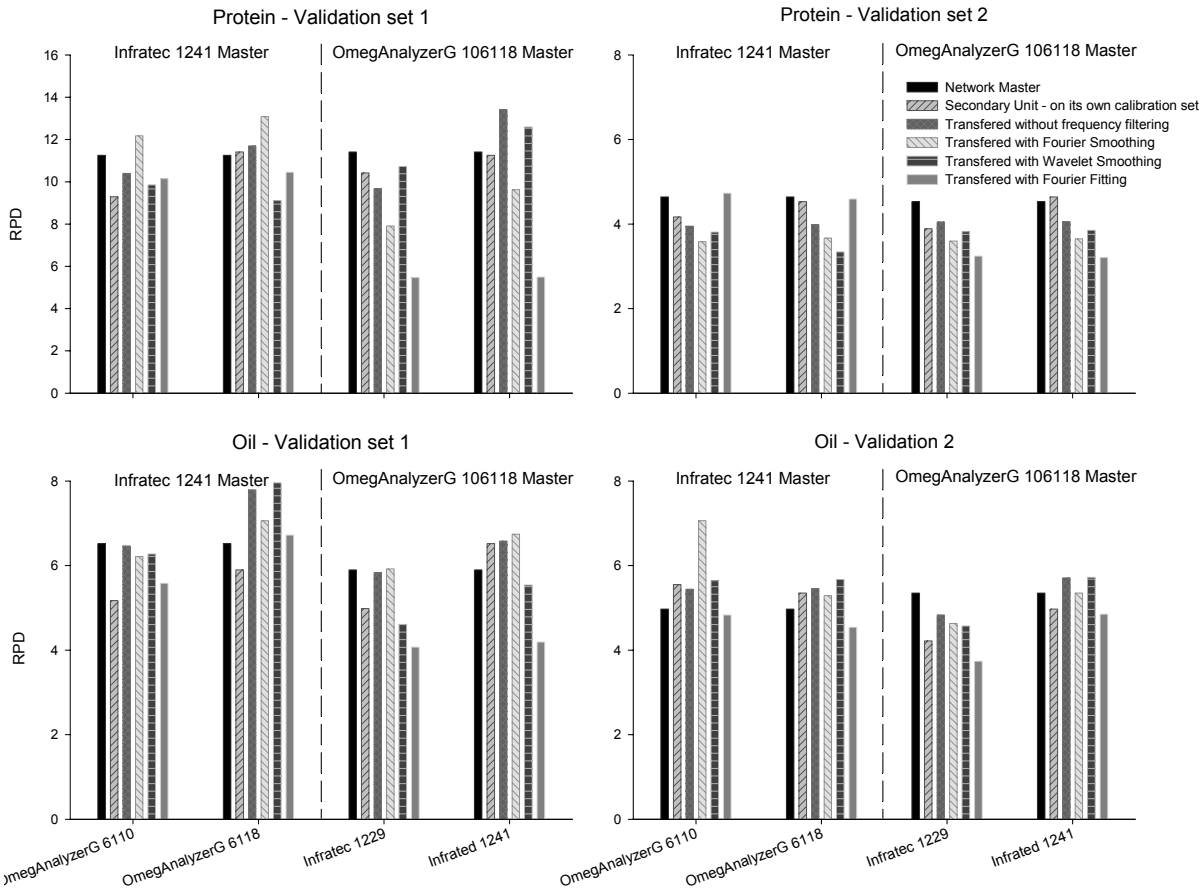


Figure 28: Evolution of secondary unit precision in inter-brand situation. Infratec 1241 was master unit of both OmegAnalyzerGs and OmegAnalyzerG 106118 was master unit of both Infratec units.

4.4.4. Effect on accuracy

When models were developed on the calibration set of each instrument and models were transferred to secondary units on the same brand, no significant differences could be observed on the bias (data not presented). All processing methods (with and without frequency modifications) had only an effect on precision, not accuracy.

However, when models were transferred across brands, there was a significant increase of bias. The loss in accuracy was significantly more important when models were

transferred from OmegAnalyzerG 106118 onto Infratecs than the opposite. Inter-brand transfer would require an additional standardization step such as a post-regression correction.

4.4.5. Conclusions

Smoothing and polynomial fitting were a good option to improve calibration precision. Similar to other common preprocessing methods, they appeared to be instrument and parameter dependent. The transfer of calibration models from instrument to instrument of the same brand and of different brands was possible. The yearly variability did not impact standardization results. Filtering in the Fourier domain gave better results, but testing with other wavelets and different filtering or fitting methods could allow the wavelet transform to perform better.

These results show that the transfer of calibrations in intra and inter-brand situations can be done by modifying the signal before calibration, for similar or better results to those obtained on the master unit. This would avoid the use of standardization methods such as optical techniques (direct standardization and piecewise direct standardization …) or post-regression corrections (slope and offset, bias correction). However, when performing inter-brand standardization with frequency filtering techniques, it will be necessary to check and correct accuracy (bias).

4.5. Robustness enhancement by sample and variable selection

4.5.1. For the development of calibration models

4.5.1.1. Local models

Figures 29.a and 29.c present validation results for both parameters and both test sets obtained with all five local methods. Bars named *Original* correspond to when all samples were used in the calibration set (results from table 8). A significant difference ($\alpha = 0.05$) could be observed between both validation sets for protein. This shows that the variability brought by next year samples impacted protein more than oil. Significant differences could be observed between Original results and all five local methods for protein and validation set 2 (all samples models performing better than selected samples). None of the sample selection methods did better than the global calibration set. This may mean that in complex scenarios, when samples to predict are different from the calibration pool, using all available information can provide the best results, no matter the amount of extra-noise included into the model.

A significant difference existed between master units and secondary units except for oil and validation set 2. This was expected because secondary unit results on Original calibrations were better than selected master unit for the OmegAnalyzerG network. This observation is of great importance regarding the choice of network master's units. If improperly chosen, the efficiency of an entire network can be jeopardized. Performances of good instruments will be lowered by a model created on a poorer dataset. Finally, Original and LOCAL performed significantly worse than other methods for oil and validation set 1 among secondary units while FDLIS and FDLIS_SA did better than all other methods for oil and validation set 2 among master units ($\alpha = 0.05$).

These results show that only in restrained situations, local methods performed better than when using all available samples. However, even though non significant, it was

possible to observe large improvements: for protein and validation set 1, LWR gave a decrease in SEP of 27% (from 0.30% pt. to 0.22% pt.); for oil and validation set 1, FDLIS gave a decrease in relative SEP of 29% (from 0.24% pt. to 0.17% pt.). When comparing precision between Original models and local methods, an improvement could be observed in 95% of the cases for protein and validation set 1, 70% of the cases for oil and validation set 1, 15% for protein and validation set 2, and 60% for oil and validation set 2.

Local chemometrics can bring a real improvement to applications where the variability to predict is known. However one must ensure that the calibration set is exhaustive enough to include the variability of new samples, since the risk taken by using local methods is higher than with global calibration models (protein and validation set 2). The variability needed to successfully predict that set was not fully included in the calibration set; local methods performed significantly worse than global models.

4.5.1.2. Feature-selected models

Figures 29.b and 29.d present validation results for protein and oil and both test sets for the use of variable selection methods. Original calibration results (all variables) were provided for comparison. Many of the observations made for local methods were true for variable selection methods as well. A significant difference could be observed between validation sets for protein. A significant difference existed between master and secondary units but for validation set 1 only. Finally, among masters, PSO performed significantly better than other methods but for protein and validation set 1 only. These results show that the use of featured selected calibration models did provided better results. Thus, in 83% of the situations, local methods did similarly or better than Original models for protein and validation set 1, 92% for oil and validation set 1, 33% for protein and validation set 2, and 92% for oil and validation set 2. The implementation of feature-selected models did provide precision improvements with a decrease of 20% relative SEP for protein and validation set 1 (from 0.3% pt. to 0.24% pt. with PSO) and of 14% for oil and validation set 1 (from 0.24% pt. to 0.21% pt. with PSO).

Selecting variables in this situation did provide an improvement, even though not significant, when predicting known variability and was not as sensitive to new variability as local chemometrics methods were. Thus, removing noise or variables not directly related to the parameter of interest can improve calibration models. The risk taken with feature-selected models is also lesser since the model is not subject to recreation based on local interferences.

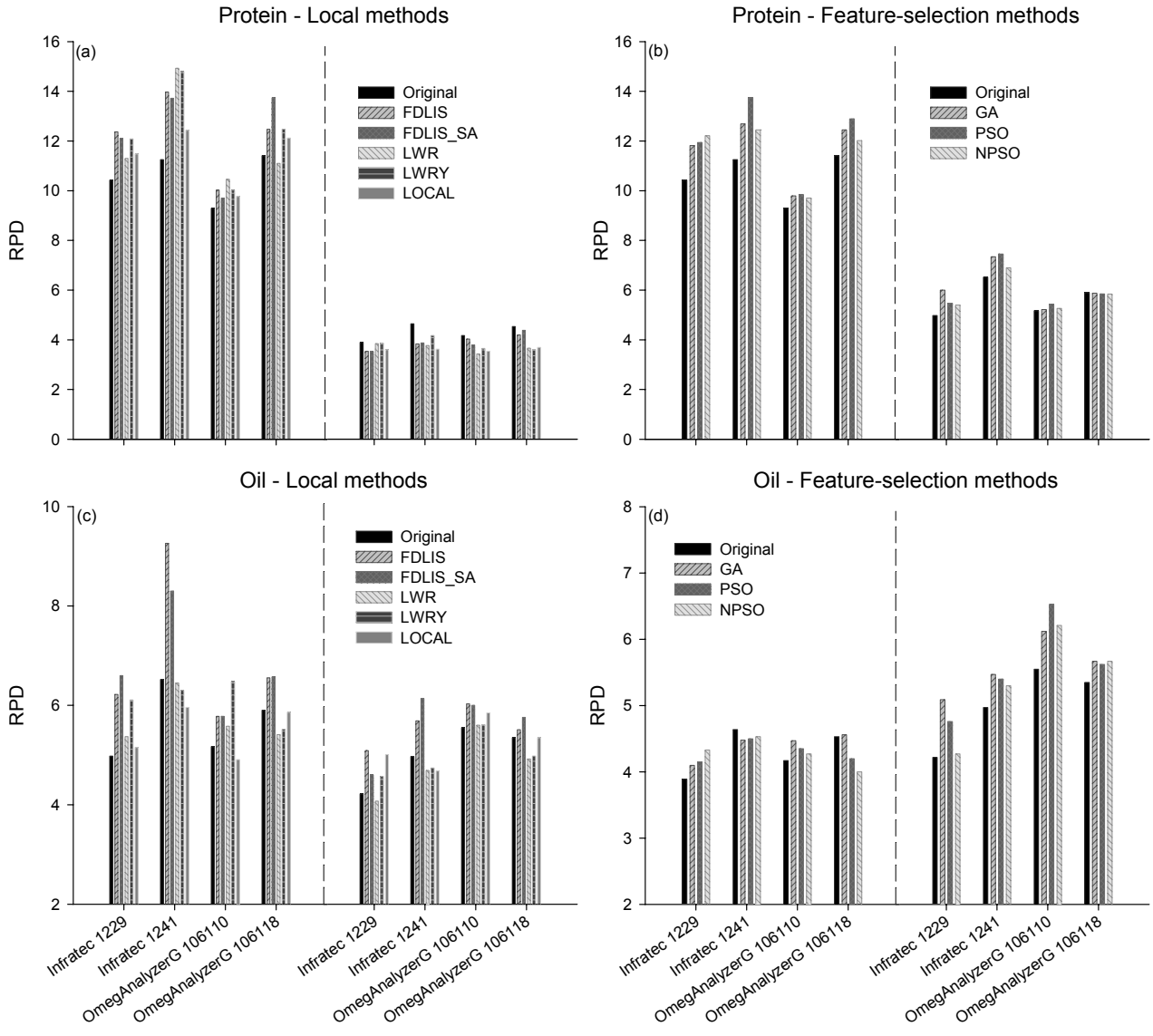


Figure 29: Validation results for local methods (top graphs) and feature-selected techniques (bottom graphs) for both parameters and validation sets when each instrument was calibrated on its own calibration set.

4.5.1.3. Local, featured-selected models

Improvement brought by local chemometrics and feature selection in certain situations motivated the development of combined models where variable selected samples were paired with local methods to predict both validation sets. All three variable selection methods were combined with all five sample selection methods. Due to the large amount of results available, only those for protein and validation set 1 are displayed as example (figure 30). The combination of methods can be powerful for precision improvement in certain situations.

Master units gave significantly better results than secondary units for protein and both validation sets. No combination performed differently on master units. However, when distinguishing between masters and secondary units, we found that among secondary units, all combinations with FDLIS and FDLIS_SA gave significantly better results for oil and validation set 1. When considering master units only, all combinations of FDLIS, FDLIS_SA, and LOCAL performed significantly better than others for oil and validation set 2.

These results show that, in certain situations, the use of these methods or a combination of methods is beneficial. For example, the combination NPSO-FDLIS_SA for protein with Infratec 1241 predicting validation set 1 provided an RPD of 17.45, a decrease in SEP of 37%. This situation, however, was not true for other parameters and validation sets.

This first set of results provided us with some very interesting information on the power of using local methods for sample and variable. Even though not generalizable to all parameters, validation sets, and instruments, a careful optimization of a calibration model can provide users with more precise results more global models. However, the risk is large and should be precisely evaluated.

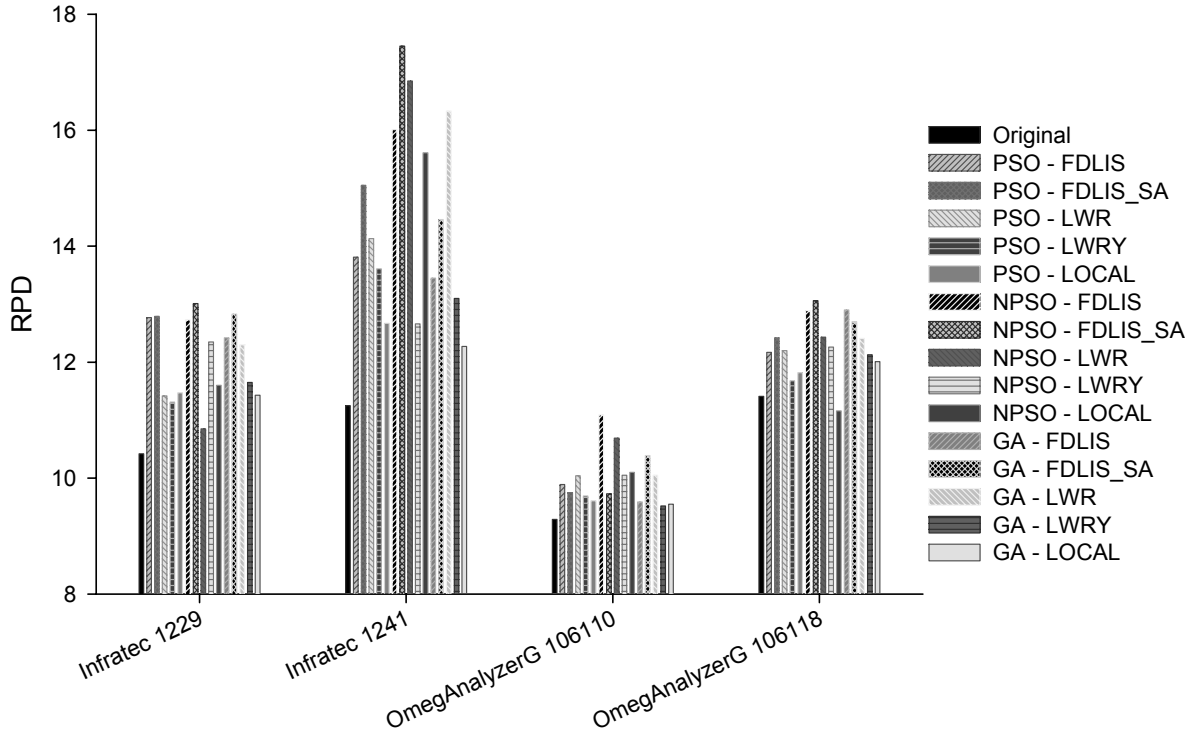


Figure 30: Validation results for combined local and feature selection methods for protein and validation set 1 only.

4.5.2. For the calibration transfer

4.5.2.1. Intra-brand situations

Figures 31 and 32 present intra-brand validation results for both validation sets and parameters for local and feature-selection methods respectively. Original performances are compared to network masters and transferred models on secondary units. For the use of local methods in calibration transfer situations (figure 31), no significant differences existed among algorithms. Thus, it is possible to transfer a local calibration model without loss of precision with these methods. Furthermore, there were no significant differences between master's results and secondary units' except for oil and validation set 1. This shows that in most situations, performances of master units were transmitted to secondary instruments. For the use of feature selection models (Figure 32), no significant differences were found

among methods. However, master and secondary units were still different on validation set 1 (both parameters). Feature selection methods could transfer calibration models across units of the same brand but not as well as local methods. No combined methods performed differently than others. A significant difference remained between prediction of master and secondary units for protein (results not shown).

These results provided the possibility to transfer calibration models developed on a master unit to secondary units of the same brand when the model has been tuned for either samples, variables or both. Local chemometrics methods appear more suitable, with a reduction of the difference between master and secondary units. Transferred local methods improved RPD compared to Original results in 60% of the cases; for variable selection techniques, improvement occurred in 79% of the situations; and for combined methods, 67% of the transferred models had a higher precision than models developed directly on secondary instruments. No trend was observable among local and variable selection methods. As in calibration, performances of the various algorithms were parameter, validation set, and instrument dependent.

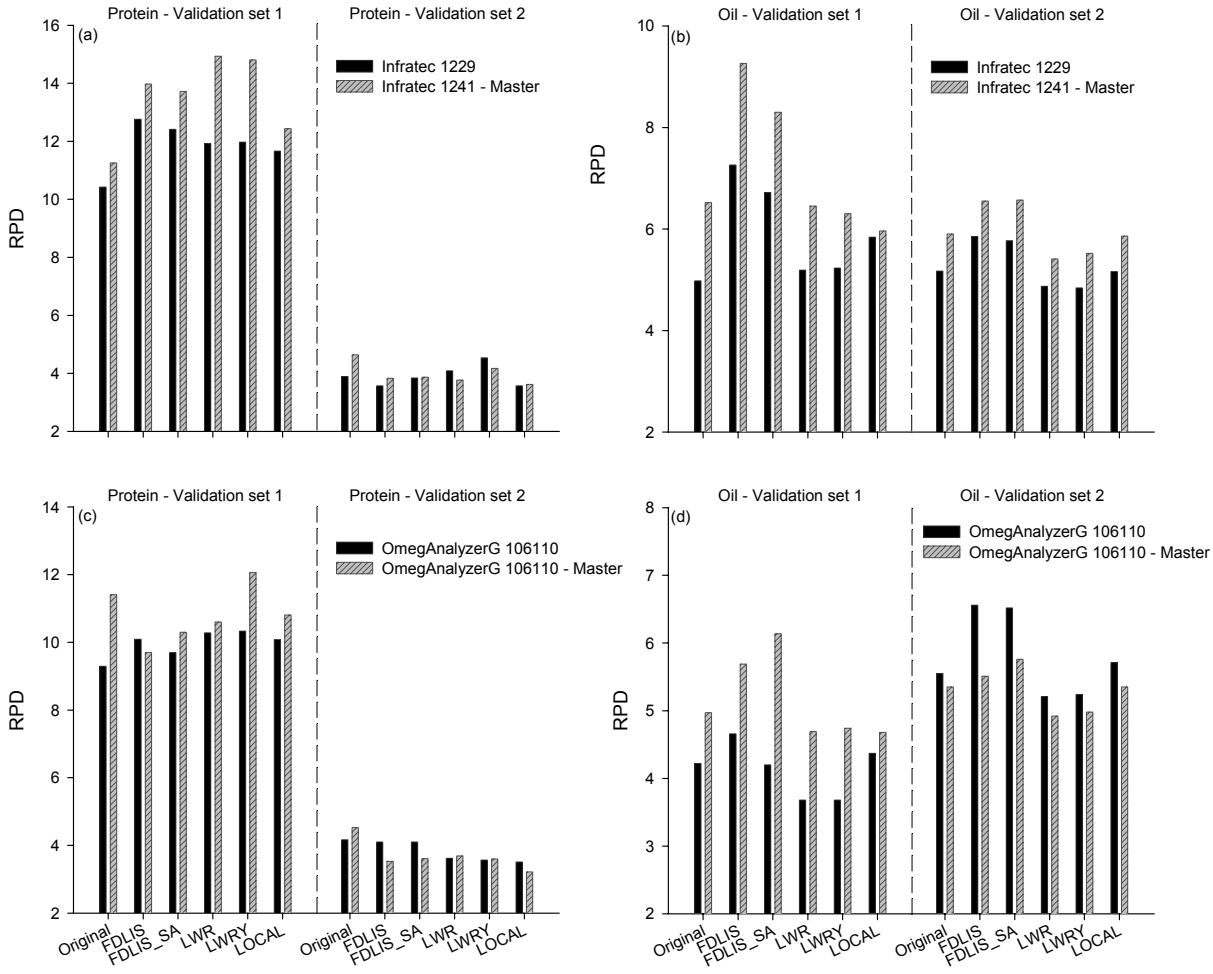


Figure 31: Validation results for local methods: both parameters and validation sets. Master units performances are compared to transferred models on secondary units of the same brand.

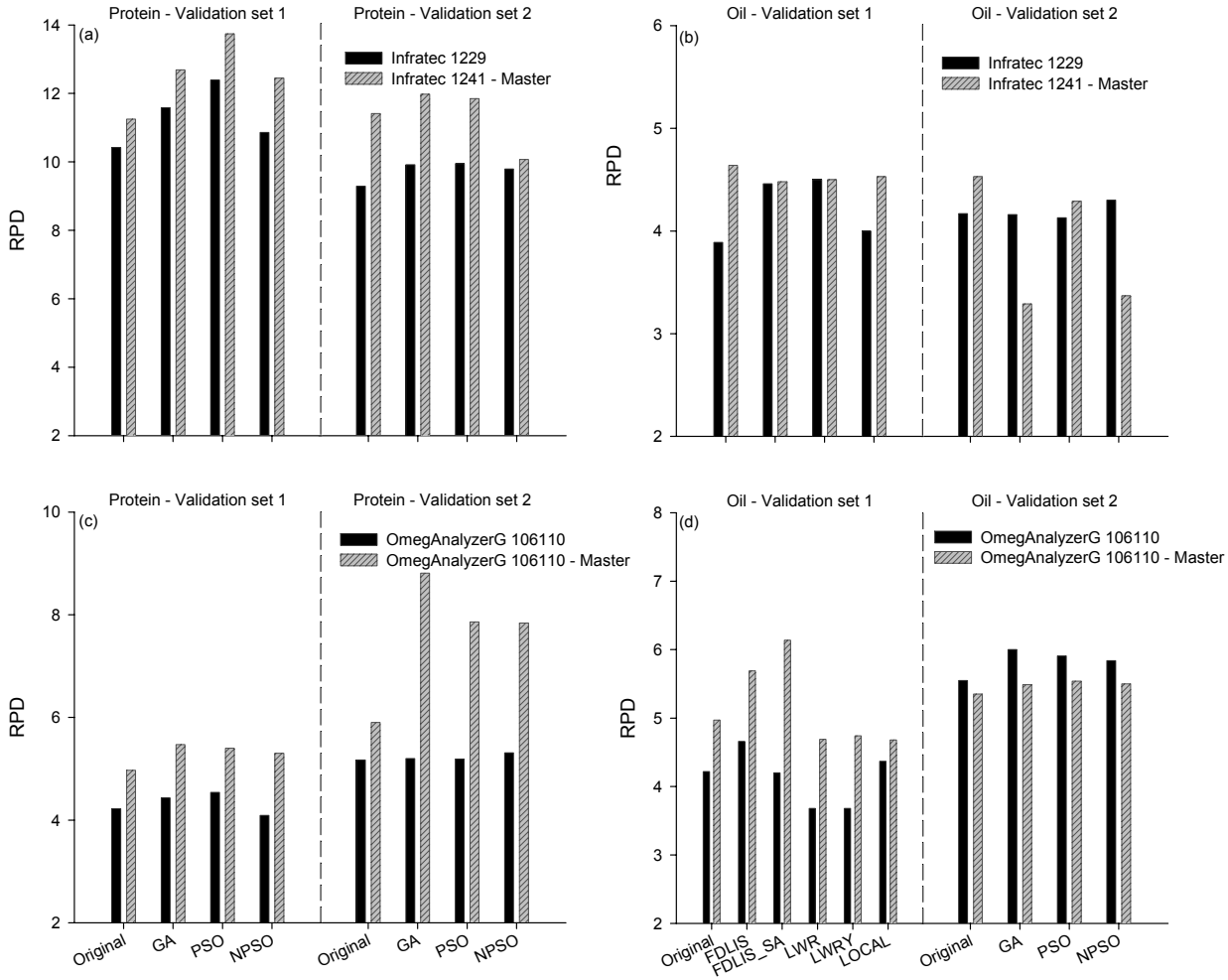


Figure 32: Validation results for feature selection methods: both parameters and validation sets. Master units performances are compared to transferred models on secondary units of the same brand.

4.5.2.2. Inter-brand situations

The ability of local methods, variable selection techniques, and their combinations to produce a model that can successfully be transferred across instrument brands was investigated: calibration models of Foss Infratec 1241 were transferred onto both Bruins instruments (figure 33) and Bruins OmegAnalyzerG models were standardized onto Infratec units (figure 34). On both figures, wavelength matching was indirectly performed by slope and bias for variable selection methods.

4.5.2.2.1. With local methods

When Foss Infratec 1241 was master of the network (figure 33.a), no significant difference could be observed between original results and transferred models performances on secondary units except for two cases. For protein and validation set 2, all local models performed poorly compared to original validation results and for oil and validation set 2, FDLIS_SA performed as well as Original and significantly better than all other techniques. In addition, validation results for secondary units and the master were not different.

When Bruins OmegAnalyzerG 106118 was master of the network (figure 34.a), FDLIS performed better than other methods for protein and validation set 1. For validation set 2, LWR did poorer than other method. LOCAL provided significantly lower results than other methods for oil and validation set 1. For three methods (LOCAL, FDLIS, and LWR), differences between master and secondary units existed for both validation set for protein and validation set 1 for oil.

The tuning of calibration model with sample selection by local chemometrics developed models transferable onto other instrument brands without loss of precision. However, it seemed that Bruins OmegAnalyzerG models were harder to transfer into Infratec units than the contrary.

4.5.2.2.2. *With feature selection methods*

The use of feature selection methods in the situation of calibration transfer where units underwent a wavelength shift required the evaluation of the impact of these non-matching variables. A simple matching system based on picking the closest variable, a slope and bias correction at the variable level (PDS window size 1), and post regression correction methods (slope and offset, bias) were evaluated in their ability to transfer feature selected models in comparison with no treatment.

When Foss Infratec 1241 was master of the network (figure 33.b), there was no difference when comparing variable selection methods, no matter the spectral matching technique used. All methods provided results similar to its own calibration set. Among matching methods, PDS performed significantly worse in all situations (parameters and validation sets). A difference between master unit and secondary units could be observed for PDS and closest variable selection method for both validation sets for protein and all variable selection methods.

When Bruins OmegAnalyzerG 106118 was master of the network (figure 34.b), PSO and NPSO performed poorly for protein and validation set 1. All methods did worse than original methods for protein and validation set 2, and GA gave lower RPD for oil and validation set 1. PDS and closest variable selection method gave significant differences between master and slaves for both validation sets for protein and all variable selection methods.

Calibration models tuned for variable selection could be successfully transferred across brands without loss of accuracy, no matter the wavelength selection method, when using no spectral matching correction and post-regression correction methods. All methods impacting the absorption values performed poorly compared to other techniques. It is also surprising that no correction of the wavelength axis gave similar results as when each model was developed on the calibration set of each instrument. This could be due to the high

collinearity among variables. Methods artificially modifying spectral trends deteriorated the relationship among variables and gave poorer results. The difficulty for OmegAnalyzerG models to be transferred onto Infratec units, which appeared with local models, was not present with variable selection methods.

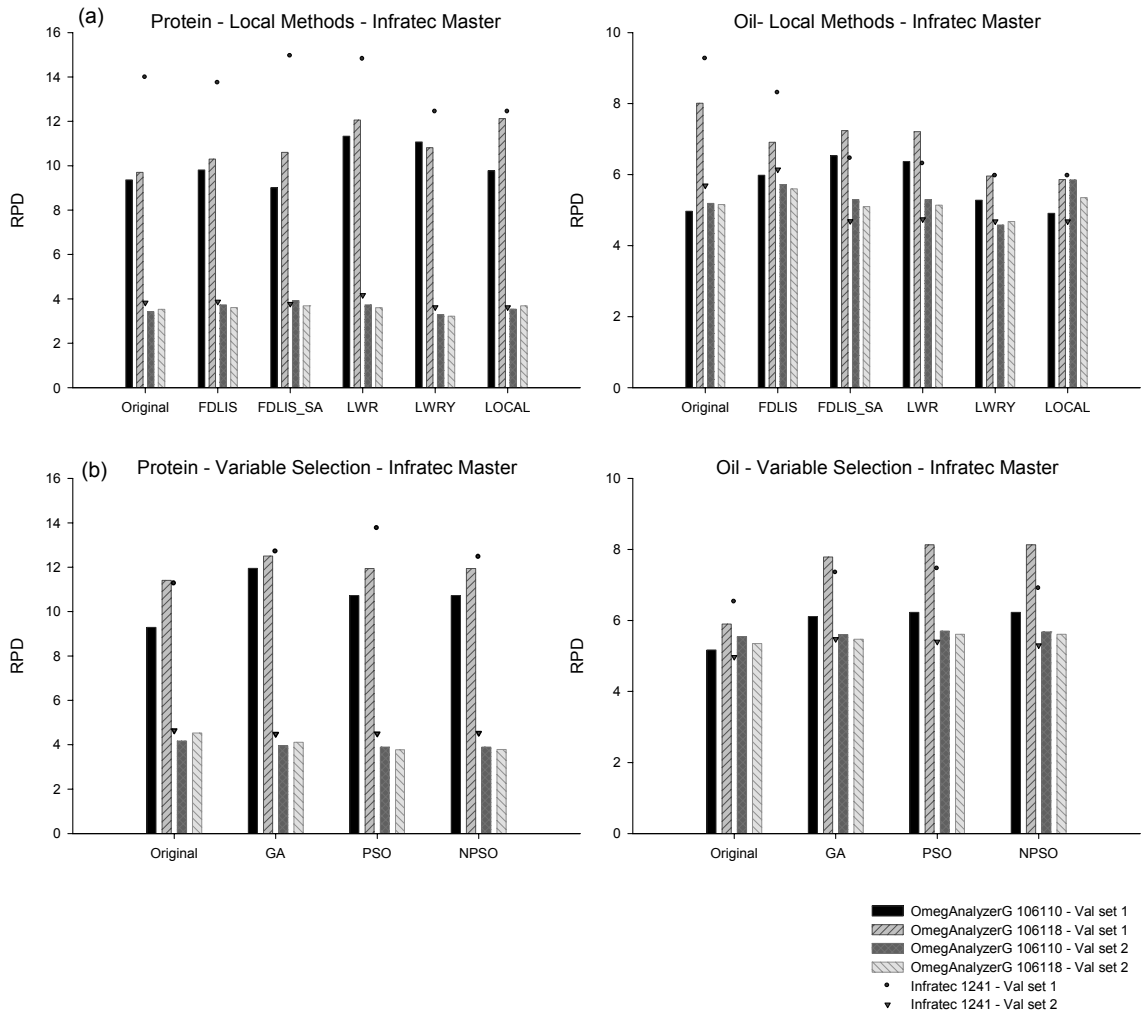


Figure 33: Validation results for local methods (top graphs) and feature-selected techniques corrected with slope and bias (bottom graphs) for both parameters and validation sets when Foss Infratec 1241 was master of both Bruins units. Bullets represent performances of master units.

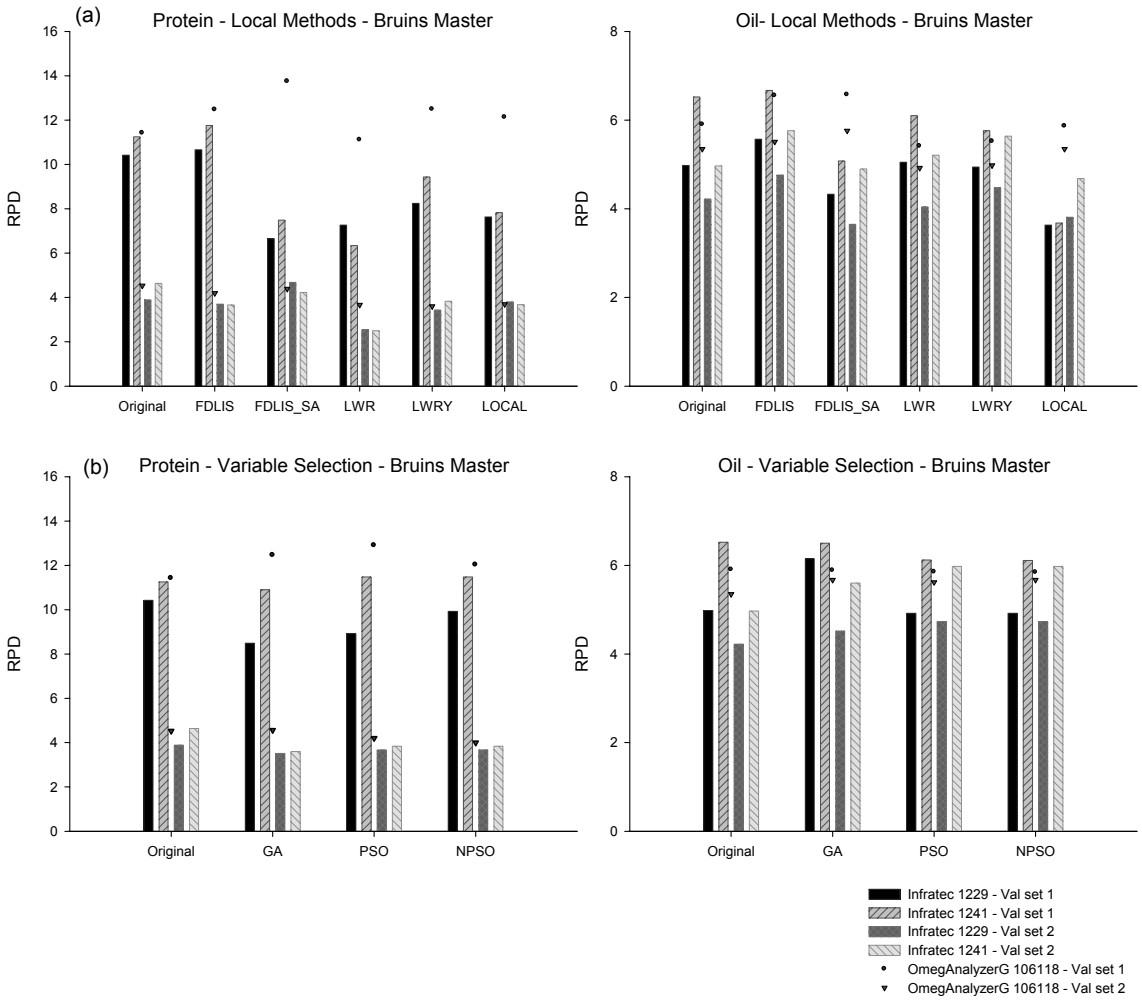


Figure 34: Validation results for local methods (top graphs) and feature-selected techniques corrected for slope and bias (bottom graphs) for both parameters and validation sets when Bruins OmegAnalyzerG 106118 was master of both Infratec units. Bullets represent performances of master units.

4.5.2.2.3. By combining local and variable selection methods

For this last set of calibration transfer results, the combinations of methods without spectral matching were compared. When Foss Infratec 1241 was master of the network, combinations PSO – FDLIS, PSO – FDLIS_SA, and PSO – LWR gave significantly higher results than other methods for protein and validation set 1. For validation set 2, NPSO – LWR only gave similar results to original calibration results, all other methods did poorer. PSO – FDLIS_SA, PSO – LWR did better than other methods for oil and validation set 2. When looking at the difference between master and slave results, four out of five combinations with GA (FDLIS_SA, LWR, LWRY, and LOCAL) were significantly different for protein (both sets), and two out of five for oil and validation set 2 (FDLIS and LOCAL).

When Bruins OmegAnalyzerG 106118 was master of the network, there were two combinations that did perform as well as original results for validation set 1 and both parameters (PSO – FDLIS, PSO – FDLIS_SA, NPSO – FDLIS, and GA - FDLIS. There was difference between master and secondary units for three of the five GA based combinations (FDLIS_SA, LWRY, and LOCAL).

4.5.2.2.4. Conclusions

Calibration transfer can successfully be done by either local methods or variable selection or both but that again, performances of the various algorithms were parameter, validation set, and instrument dependent. Spectral differences between brands were easily overcome by the various methods, but no clear winner could be determined. Among variable selection methods, it was harder to transfer calibration models from OmegAnalyzerGs to Infratecs than the contrary.

4.5.3. Evaluation of selected variables

Variable selection methods provided good results in calibration and in intra-brand calibration transfer situations. However, they showed limitations in inter-brand situations. Figure 35 presents selected variables from all three methods for protein models. Among selected variables, there were 45% common between PSO and NPSO, 44% between NPSO and GA, and 34% between PSO and GA. There was no significant difference among all three methods even though they were based on three very different variable patterns. It appears that all main chemical families are present in each selections (CHs, H_2O, ROH, and RNH_2), but not with the same relative importance. Each algorithm was able to select a set of variable that "worked well", but that did not necessarily constitute the optimal selection.

Local methods may not be the most appropriate methods to consider with the type of data used in this study. The chance to find local optima is high with instability of the final model.

4.5.4. Model accuracy

Model accuracy (bias) is also of importance when considering the use of calibration and standardization methods. In this study, all methods tested did not have a significant effect on accuracy. They all performed as well as original calibrations.

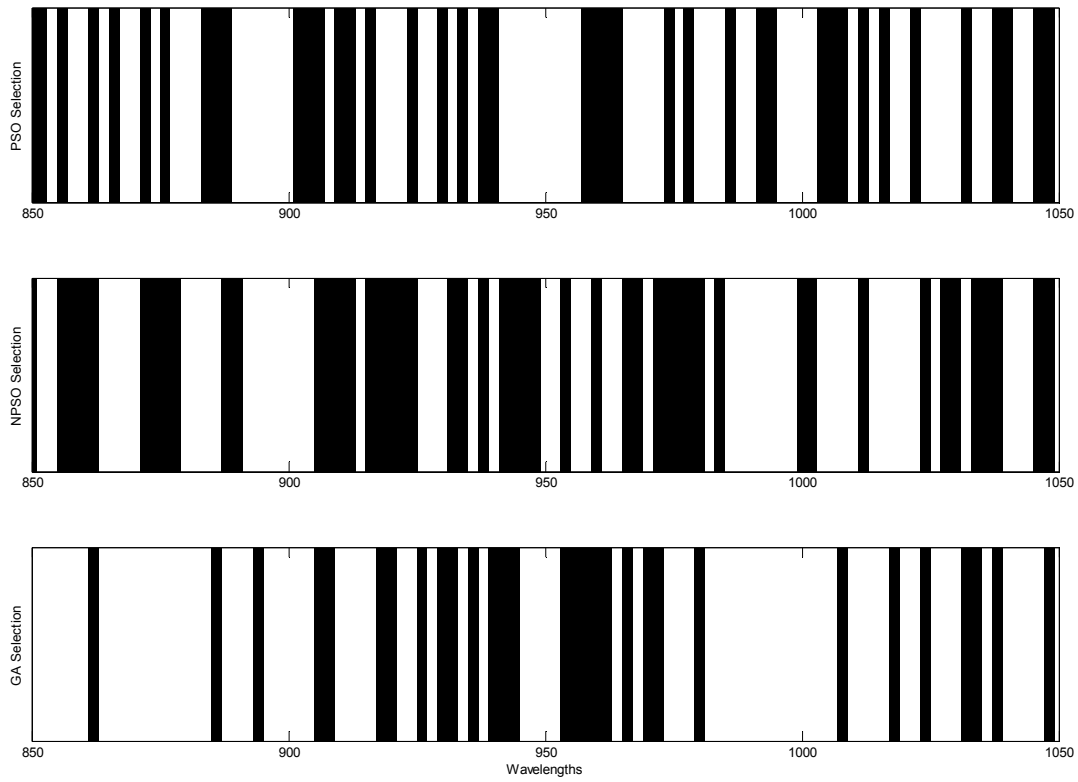


Figure 35: Variables selected by the three variable selection methods implemented in this study (PSO: particle swarm optimization, NPSO: neighboring particle swarm optimization, and GA: genetic algorithms). Selected variables are represented in black.

4.5.5. Concluding remarks

In this section, we evaluated the ability of local chemometrics methods, variable selection techniques, and their combination for calibration development and transfer in intra and inter-brand situations. Even though no clear trends could be found, all methods provided, in certain circumstances, precision improvements in calibration or standardization or both. Optimization methods were instrument, parameter, and situation dependent. Local and variable selection algorithms can be used for calibration and calibration transfer, but it is necessary to evaluate the variability of future samples. FDLIS

and FDLIS_SA were as competitive local methods versus existing algorithms, and NPSO performed similarly to PSO and GA.

Risks taken must be evaluated when using these methods. Local chemometrics for sample and variable methods are attractive, but can be the source of instability and erroneous results if performed in uncontrolled situations. The robustness they bring is relative to the variability within the calibration set, due to the specialization they create.

4.6. Orthogonal based robustness enhancement methods

4.6.1. Intra-brand calibration transfer

Figure 36 presents, for each parameter and brand network, validation results on both validation sets. Performances of models developed on master units and transferred onto secondary unit are compared with their respective master units. Dotted lines serve as reference for master results since they were unchanged (and were omitted on the figure) by all orthogonal methods except O-PLS and OSC.

Validation sets were different for protein ($\alpha = 0.05$) but not for oil. The limited range for protein in validation set 2 was most likely responsible for these results along with new variability brought in by 2006 crop year.

For protein, TOP and EROS gave significantly lower RPDs than other techniques for validation set 1 (true for both brands). The seven other methods did not give different results. For the validation set 2, there was no difference. For oil and both validation sets, there were also no significant differences. Even though results were not significant, not all methods behaved the same way. Tables 10 and 11 summarize the performance of the methods. Many of them gave better precisions than original models (table 10). A couple of situations provided secondary units with better performances than the master unit (table 11). This is an indication that removing the uninformative space representing the difference between instruments can have a beneficial effect. This also shows that not all methods are appropriate for all parameters and samples. A strong validation strategy must confirm the appropriateness of the method to the situation of interest. It is interesting to notice that TOP 4years and EROS 4years did not perform better than other methods even though they were designed to remove the variability through years. Sequential and block DOPs performed well, especially for the Bruins network, while DOP did not do well for that same network. The inclusion of more variability into the difference matrix helped to remove the external information. Validation set 1 was less well predicted than validation set 2. This could be

explained by the fact that, since the validation set 1 and the calibration set are very close, too much informative dimensions were removed, limiting the predictive ability of the model. However, the model robustness was enhanced when predicting samples different from those in the calibration set. This was particularly true for oil. This could mean that orthogonal methods trend to remove informative information causing model to underfitting the data. This can be beneficial or a significant constraint depending on the application.

In the present situation, TOP and EROS, as well as TOP 4years and EROS 4years respectively gave the same results. Figure 37 presents the first four loadings of the $\boldsymbol{D}$ matrix between Infratec 1229 and Infratec 1241, after a PCA decomposition obtained from TOP (figure 37.a) and EROS (figure 37.b). Both methods provided the same correction space. In TOP, the internal difference within classes (a class being a sample collected on several instruments) is used to form a matrix whose variance is analyzed while in EROS, the variance for each class is calculated, summed, and that sum is analyzed. In situations where classes are the same, TOP and EROS provide the same results. TOP/EROS and TOP/EROS 4years will be used from now and on to represent these methods.

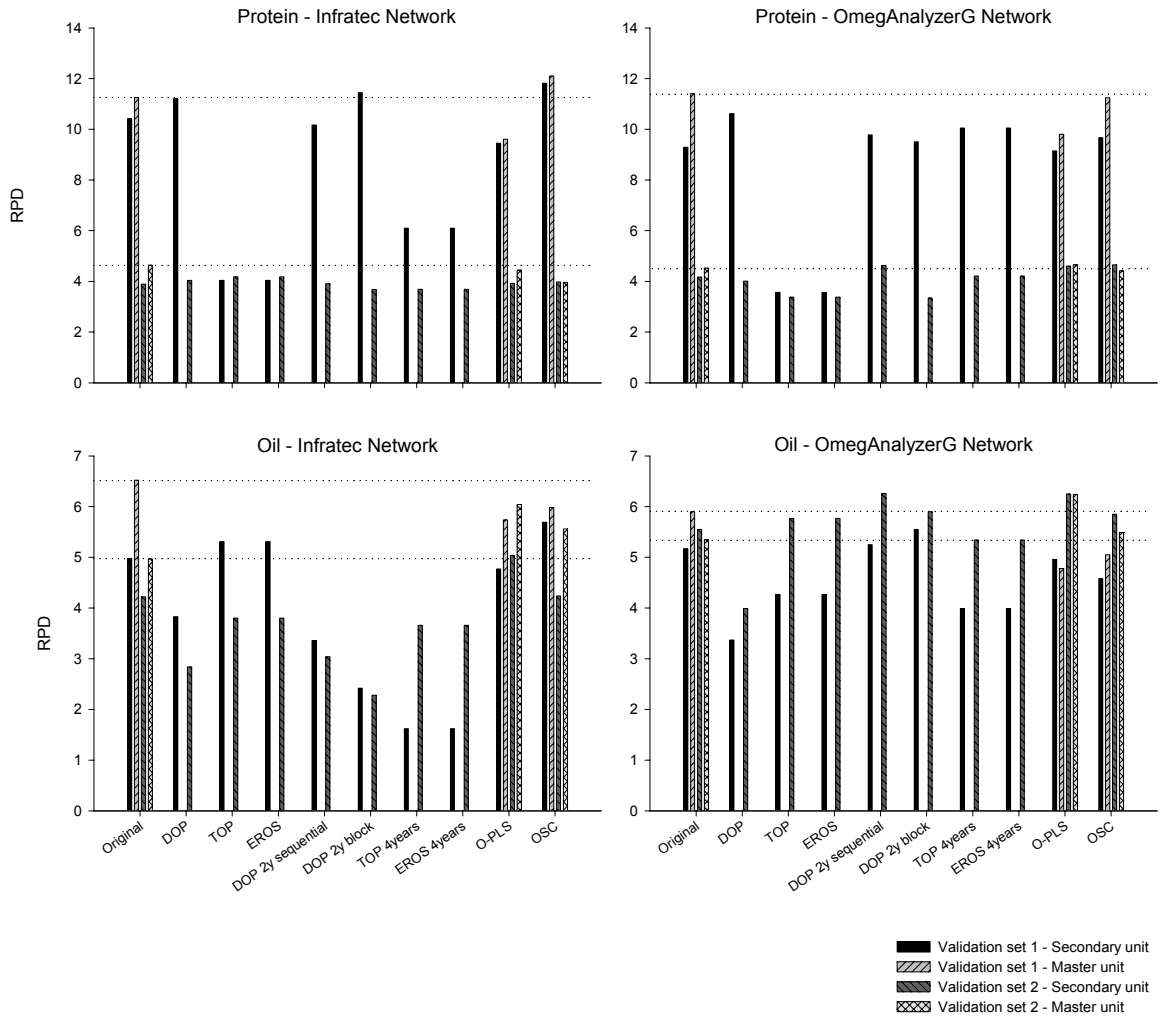


Figure 36: Validation of the standardization methods in an intra-brand situation (Infratec 1241 is the master unit for Infratec 1229; OmegAnalyzerG (s/n: 106118) is the master unit for OmegAnalyzerG (s/n: 106110)). Dotted lines correspond to benchmark results.

Table 10: Precision gain or loss provided by each standardization method compared to original calibrations.

		Protein		Oil		Number of latent variables		Number of component removed	
Method	Instrument	Val. set 1	Val. Set 2	Val. set 1	Val. Set 2	Protein	Oil	Protein	Oil
DOP									
	Infratec 1229	+	+	-	-	10	12	1	1
	OmegAnalyzerG 106110	+	-	-	-	11	9	1	3
TOP									
	Infratec 1229	-*	+	+	-	12	11	1	1
	OmegAnalyzerG 106110	-*	-	-	+	12	11	1	1
EROS									
	Infratec 1229	-*	+	+	-	12	11	1	2
	OmegAnalyzerG 106110	-*	-	-	+	12	11	1	1
DOP 2y sequential									
	Infratec 1229	-	=	-	-	12	12	1	1
	OmegAnalyzerG 106110	+	+	+	+	12	12	1	1
DOP 2y block									
	Infratec 1229	+	-	-	-	11	12	1	1
	OmegAnalyzerG 106110	+	-	+	+	11	11	1	1
TOP 4years									
	Infratec 1229	-	-	-	-	9	10	2	3
	OmegAnalyzerG 106110	+	=	-	-	10	11	1	3
EROS 4years									
	Infratec 1229	-	-	-	-	9	10	2	3
	OmegAnalyzerG 106110	+	=	-	-	10	11	1	3
O-PLS									
	Infratec 1229	-	=	-	+	12	9	2	1
	OmegAnalyzerG 106110	-	+	-	+	12	8	2	3
OSC									
	Infratec 1229	+	+	+	=	10	9	1	1
	OmegAnalyzerG 106110	+	+	-	+	11	12	1	1

+ result of the transferred secondary unit greater than the original secondary unit

- result of the transferred secondary unit lower than the original secondary unit

= results equivalent between transferred and original models

* marks significance

Table 11. Precision gain or loss provided by each standardization method compared to network masters.

Method	Instrument	Protein		Oil	
		Val. set 1	Val. Set 2	Val. set 1	Val. Set 2
DOP					
	Infratec 1229	=	-	-	-
	OmegAnalyzerG 106110	-	-	-	-
TOP					
	Infratec 1229	-	-	-	-
	OmegAnalyzerG 106110	-	-	-	+
EROS					
	Infratec 1229	-	-	-	-
	OmegAnalyzerG 106110	-	-	-	+
DOP 2y sequential					
	Infratec 1229	-	-	-	-
	OmegAnalyzerG 106110	-	+	-	+
DOP 2y block					
	Infratec 1229	+	-	-	-
	OmegAnalyzerG 106110	-	-	-	+
TOP 4years					
	Infratec 1229	-	-	-	-
	OmegAnalyzerG 106110	-	-	-	-
EROS 4years					
	Infratec 1229	-	-	-	-
	OmegAnalyzerG 106110	-	-	-	-
O-PLS					
	Infratec 1229	-	-	-	+
	OmegAnalyzerG 106110	-	=	-	+
OSC					
	Infratec 1229	-	-	-	-
	OmegAnalyzerG 106110	-	=	-	+

\+ result of the transferred secondary unit greater than the master unit

\- result of the transferred secondary unit lower than the master unit

= results equivalent between transferred and master models

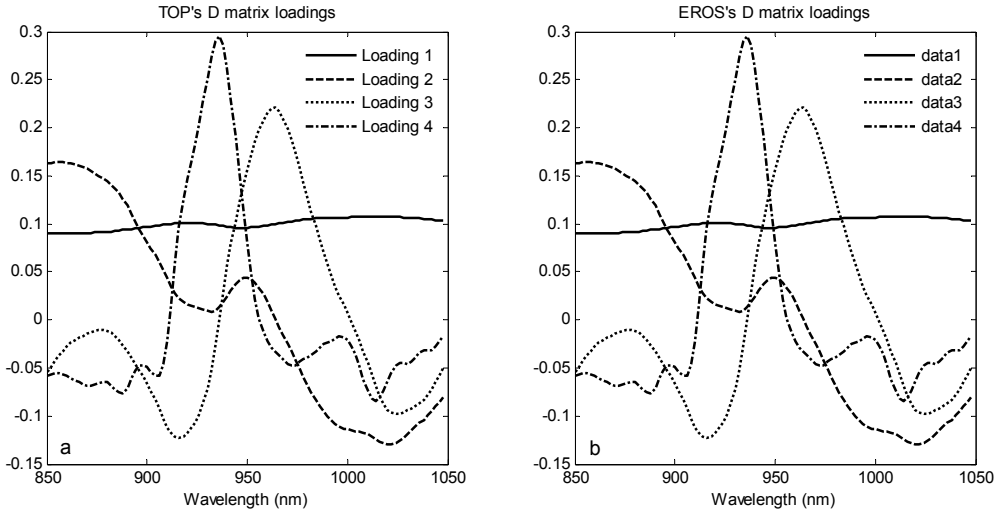


Figure 37: Loading vectors of the difference matrix ***D*** between Infratec 1229 and Infratec 1241 from a TOP decomposition (figure 37.a) and an EROS decomposition (figure 37.b).

4.6.2. Inter-brand calibration transfer

Figures 38 and 39 respectively present, for each parameter and validation set, the performances of an Infratec model to be transferred onto Bruins units and vice-versa. For both situations (Infratec 1241 and OmegAnalyzerG master of the global network), and for both parameters, there is always a method that will provide satisfactory calibration transfer results. The transfer of calibration models across brands is possible and provides results as good as or better than if models were developed on their own calibration sets.

For Foss Infratec 1241 master of the Bruins OmegAnalyzerG network, there were no significant results for protein and validation set 1, but DOP, DOP 2y block, and OSC performed the best. For validation set 2, only TOP/EROS provided as good results as the original model with all six other methods significantly less precise ($\alpha = 0.05$). For oil and validation set 1, TOP/EROS, DOP 2y sequential and DOP block provided significantly lower RPDs than other methods while TOP/EROS 4years, and OSC provided OmegAnalyzerG 106118 with significantly higher results ($\alpha = 0.05$). Finally, DOP 2y block performed again less precisely than other orthogonal methods ($\alpha = 0.05$).

For Bruins OmegAnalyzerG 106118 master of the Foss Infratec network (Figure 4), there was no significant difference for both parameters and validation sets. Depending on the situations, different methods performed better than others, but this was not consistent from parameter to parameter and validation set to validation set.

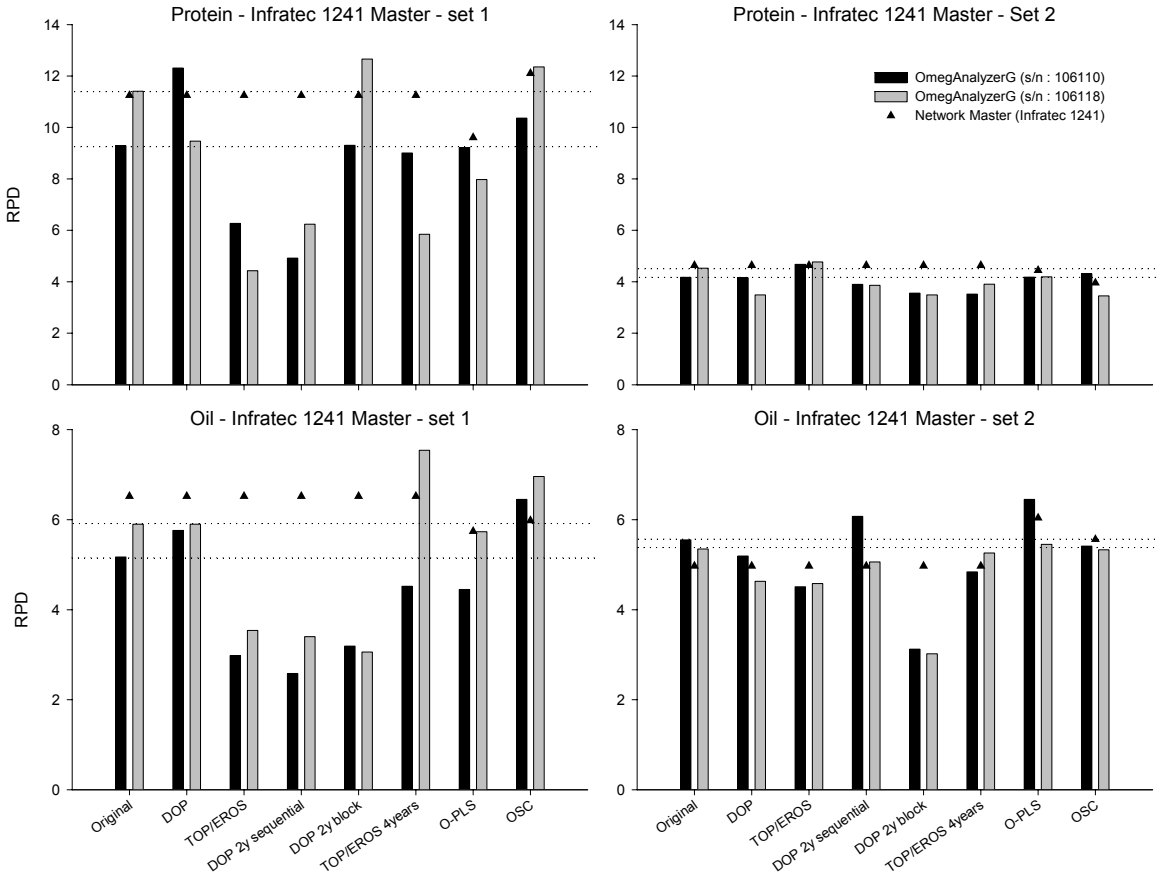


Figure 38: Validation of the standardization methods in an inter-brand situation with Foss Infratec 1241 master of the network. Bars represent secondary units while triangles are the performances of the master unit. Dotted lines correspond to benchmark results.

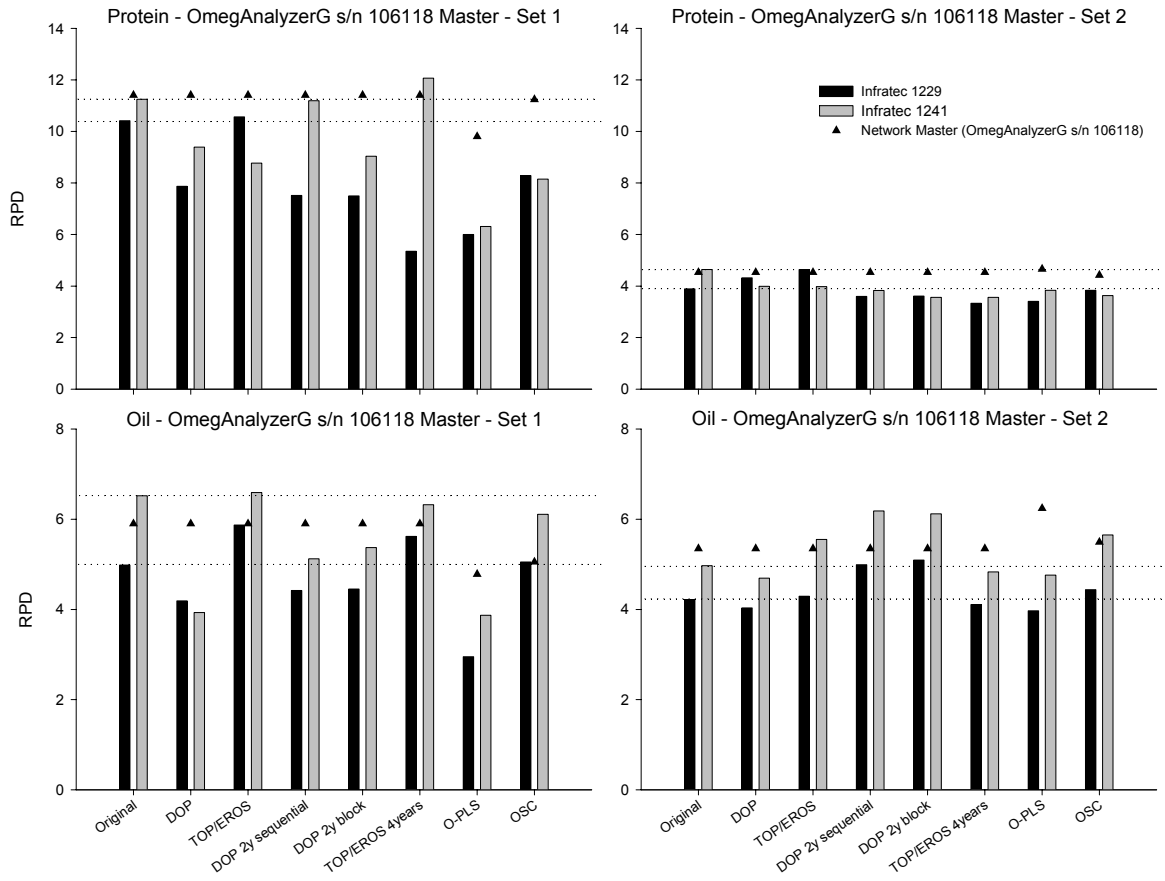


Figure 39: Validation of the standardization methods in an inter-brand situation with Bruins OmegAnalyzerG s/n 106118 master of the network. Bars represent secondary units while triangles are the performances of the master unit. Dotted lines correspond to benchmark results.

4.6.3. Evolution of the accuracy

Improving or maintaining model precision is only part of the calibration transfer work. Keeping the accuracy (mean difference between predicted and reference values) as good as on the master unit is also challenging. In this study, intra-brand calibration transfer scenarios did not impact accuracy. Models performed as accurately on the master unit as on the secondary units. However, in inter-brand scenarios, orthogonal methods improved model accuracy which was not the case for exhaustive calibration methods and slope and bias ones (data not shown). This may signify that a two-step calibration transfer process may be needed: one to maintain precision and a second to correct for bias loss.

4.6.4. Conclusions

This section compared the use of orthogonal methods with other common standardization techniques for the transfer of calibration models between instrument of the same brand and across brands. Orthogonal methods were shown to perform similarly to secondary units when calibrated on their own calibration sets. While no method outperformed others, no beneficial effect was observed when using a collection of years representing instrumental variability present in the calibration set (TOP/EROS 4years). Modifications to DOP provided good results and orthogonal methods using information in the Y matrix (O-PLS and OSC) performed better than most of other methods. However, the later require the transformation of the validation samples and may not be suitable for all instrumental setups.

The transfer of calibration models from brand to brand was shown possible, with an improvement of initial model precision. A control of model fidelity must be performed to ensure proper predictions. The two validation strategies finally showed that it is necessary to evaluate the transferred models with the variability that it will encounter, different transfer methods performing better on different types of spectral variations.

CHAPTER 5.

GENERAL DISCUSSION

At the end of chapter two, a quote from Anthony M.C. Davis reminded the readers that the implementation of standardization methods is only beneficial if an improvement is observed. A successful calibration transfer can either mean that prediction performances of the secondary units are lower, but within specific limits, similar or better than those on the master unit, or that prediction performances of the secondary units are similar or better than when models are developed directly with secondary unit spectra. The choice of the master unit is consequently of importance. A wrong choice of this unit can bring a decrease of overall network performance since a low quality model will not perform as well as a high quality model could on lower quality instruments.

AACC method 39.00 states that the standardization error should not be larger than 10% of the original precision. Throughout this study we have seen that in many cases (depending on the method used, the instrument, and the validation set), there was no significant difference between master and secondary unit predictions ($\alpha = 0.05$). Also, in many situations, the comparison between benchmark results and standardized results showed that most methods gave similar or better results than when each instrument was calibrated on its own calibration set.

Chapter 2, 3, and 4 presented, developed, compared, and applied fifty-seven calibration transfer methods of which thirty-five could be used in calibration development (these methods modified the calibration set and not only the prediction results on the secondary units). For each family of methods (common techniques, high frequency components filtering, sample and variable selection techniques, and orthogonal methods), there was no significant difference between benchmark results and transferred models for most of the methods except for two common techniques (DS and PDS). However,

performances themselves were not compared. Table 12 presents for the two secondary units in intra and inter brand situations, the percentage of cases where the 10% decrease in SEP rule was met. For all fifty-seven calibration transfer scenarios, the SEP obtained after transfer was compared with network masters and benchmarks. Large variability existed among methods, which was impacted by validation sets, parameters, and instruments. Satisfactory results were obtained when compared with benchmark, but the comparison with the network master showed that in some situations, the 10% rule was met by only very few methods.

Table 12: Percentage of cases where the 10% decrease in SEP was met (includes all 57 methods).

		Intra-brand scenario		Inter-brand scenario	
Instrument	Validation set	Compared to network master	Compared to benchmark results	Compared to network master	Compared to benchmark results
Infratec 1229	Protein 1	46%	91%	7%	25%
	Protein 2	51%	86%	33%	74%
	Oil 1	30%	91%	46%	72%
	Oil 2	5%	82%	5%	81%
OmegAnalyzerG 106110	Protein 1	11%	89%	16%	84%
	Protein 2	86%	70%	39%	51%
	Oil 1	33%	88%	16%	81%
	Oil 2	95%	91%	82%	82%

To further detail these results, changes in precision were presented and compared for all methods used in this study. Figures 40 to 49 present calibration and standardization results for intra and inter-brand scenarios. These figures use RPD and not SEP. But since the relation between both parameters is dependent on the standard deviation of the validation set, observations made with RPD are comparable to those made with SEP.

Figures 40 and 41 present the results of the thirty-five methods that could be used in calibration. For protein and Infratec 1229, precision was equal or better in 97% of the cases for validation set 1 and in 66% of the cases for validation set 2. For OmegAnalyzerG 106110, RPD was increased or remained constant in 91% and 49% of the cases for validation set 1

and 2 respectively. The impact of the shorter range and lower standard deviation of validation set 2 was greater for OmegAnalyzerG 106110 than for Infratec 1229. Differences mainly occurred when local methods were used (with or without variable selection methods). It is difficult to explain these differences, but it is possible to say that the way "local" samples are selected may not be completely appropriate in cases of validation sets with a sample distribution different from the one of the calibration set. More work on creating a homogeneous database should be done. For oil, while Infratec 1229 gave equal or higher precision in 97% of the cases for validation set 1, it was the case of only 66% of the situations for OmegAnalyzerG 106110. Globally, OmegAlyzerG 106110 did not react well to frequency filtered spectra and orthogonal methods. For validation set 2, the situation was inversed since Infratec 1229 did well in 83% of the cases while OmegAnalyzerG 106110 performed better in 91% of the scenarios. This is certainly due to the removal of too much information when filtering noise.

It is difficult to draw conclusions from such variable results. There was not a method that performed worse in all situations (instrument, parameter, validation set), but also none that performed better. However, in specific situations, large decreases in prediction error occurred. Chemometricians should be aware of these variations and address various experimental situations with the appropriate method or combinations of methods through an iterative optimization process.

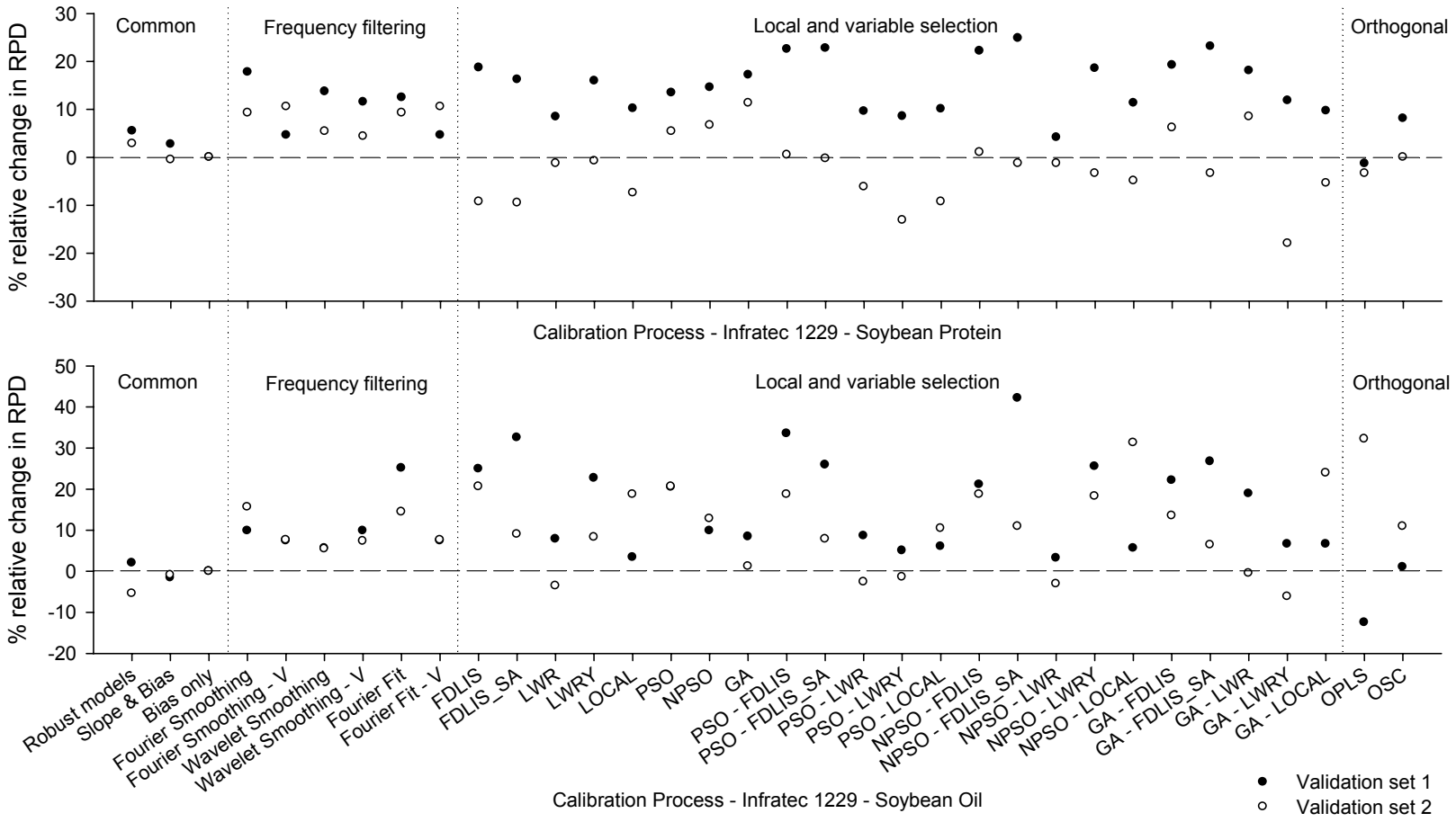


Figure 40: Comparison of calibration process for Infratec 1229 (using Infratec 1229 calibration and validation sets). V stands for when validation sets were preprocessed.

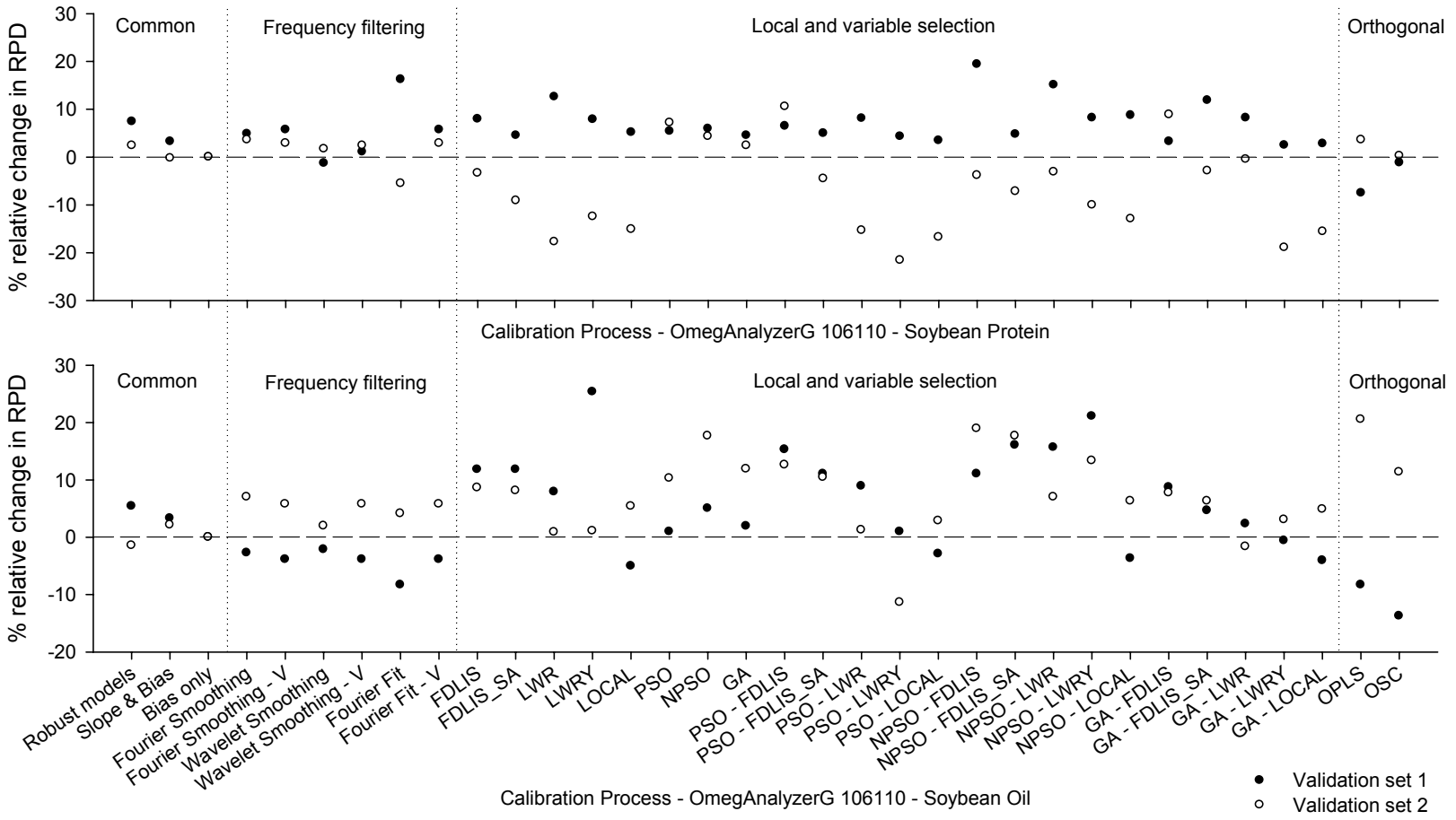


Figure 41: Comparison of calibration process for OmegAnalyzerG 106110 (using OmegAnalyzerG 106110 calibration and validation sets). V stands for when validation sets were preprocessed.

The next four figures compare prediction precisions for all fifty-seven methods when each instrument was calibrated on its own calibration set for intra-brands (figures 42 and 43) and inter-brand (figures 44 and 45) calibration transfer scenarios.

In intra-brand situations, the comparison of each method with the benchmark showed that filtering methods along with most local and variable selection methods performed well for both protein and oil for Infratec 1229. The use of y based orthogonal methods gave better results than **X** only based methods. For OmegAnalyzerG 106118, improvements were not as large and some combinations of local and variable selection methods did not perform well for protein. For oil, frequency based methods as well as local and variable selection technique improved benchmark results.

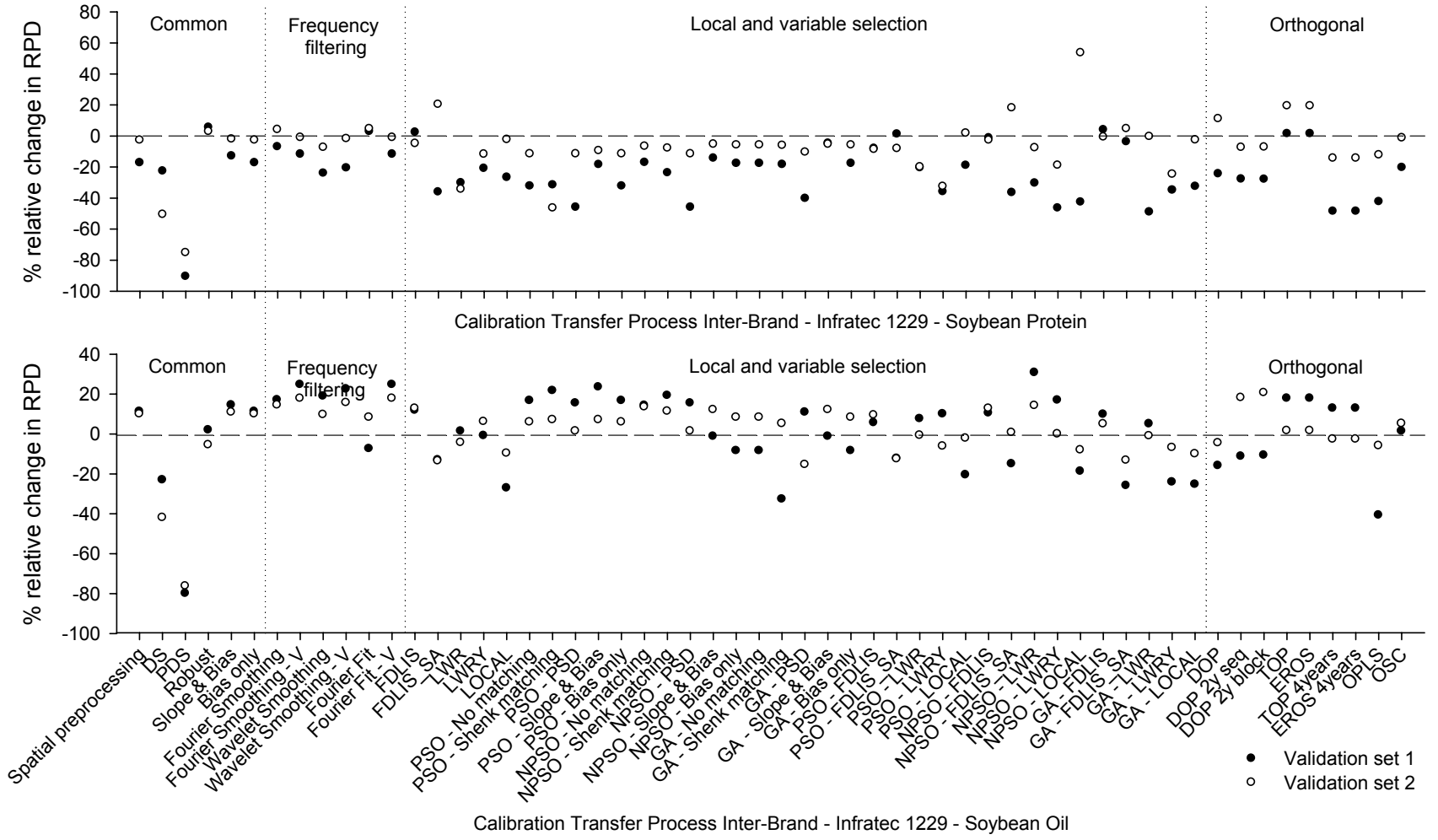


Figure 42: Comparison of intra-brand calibration transfer for Infratec 1229 (Infratec 1241 master) with original calibration results. V stands for when validation sets were preprocessed.

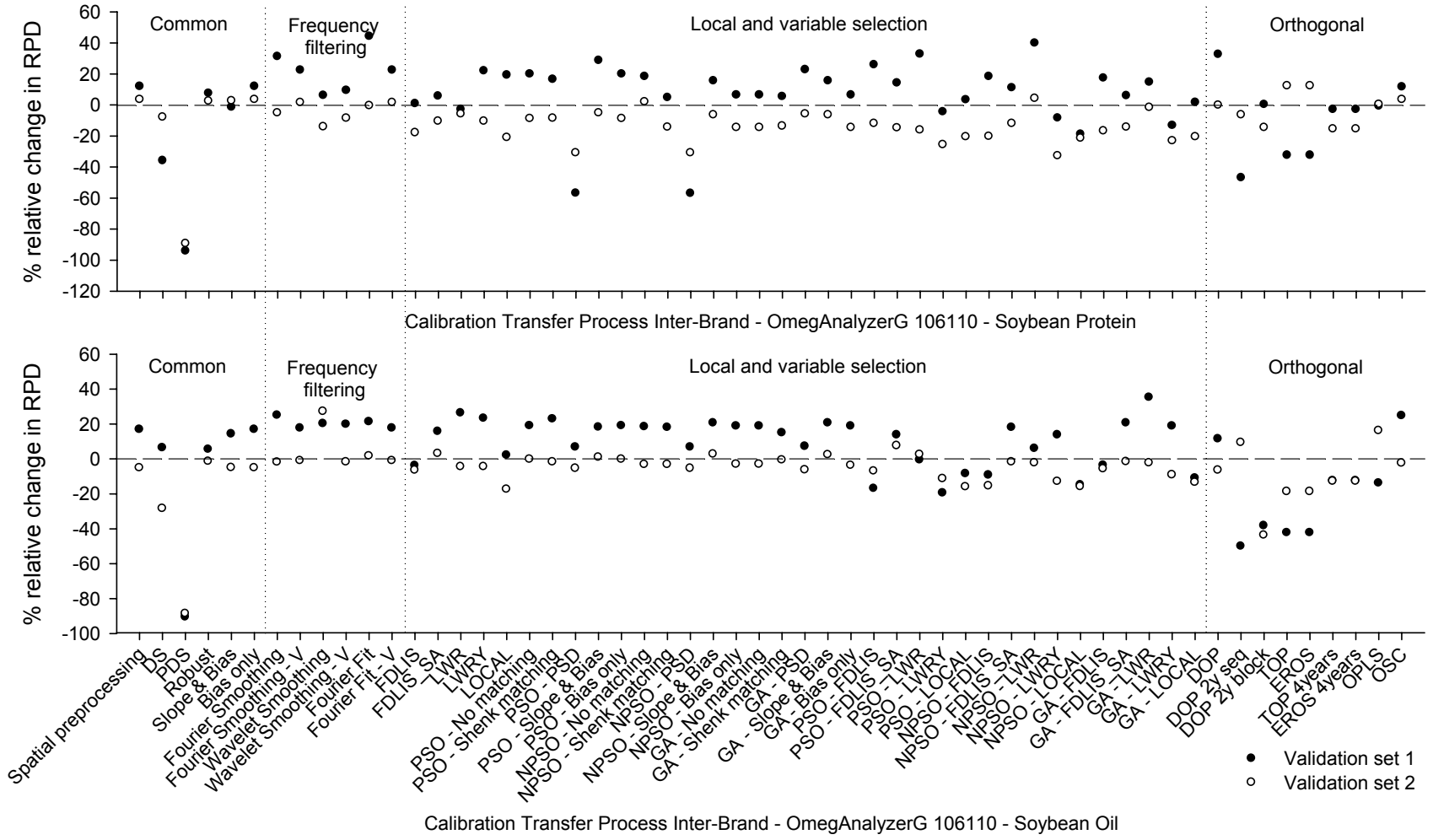


Figure 43: Comparison of intra-brand calibration transfer for OmegAnalyzerG 106110 (OmegAnalyzerG 106118 master) with original calibration results. V stands for when validation sets were preprocessed.

In inter-brand situations, similar observations to intra-brand calibration transfer could be made. Except for protein and Infratec 1229, where almost all methods performed poorly compared to benchmark results, trends observed on figure 42 and 43 were confirmed in inter-brand standardization. Frequency and sample/feature selection based methods performed well. Results from orthogonal methods varied with instruments and parameters.

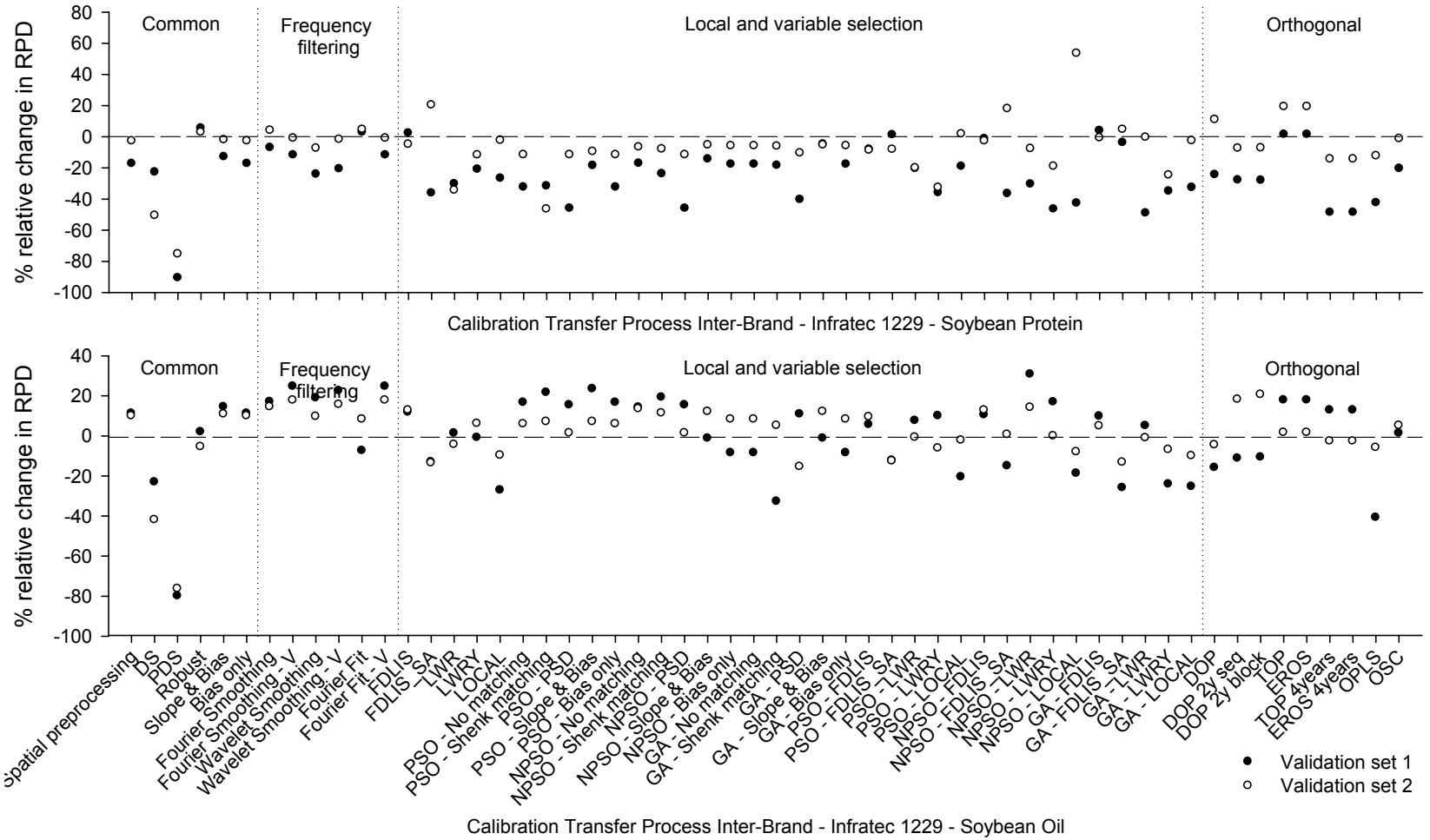


Figure 44: Comparison of inter-brand calibration transfer for Infratec 1229 (OmegAnalyzerG 106118 master) with original calibration results. V stands for when validation sets were preprocessed.

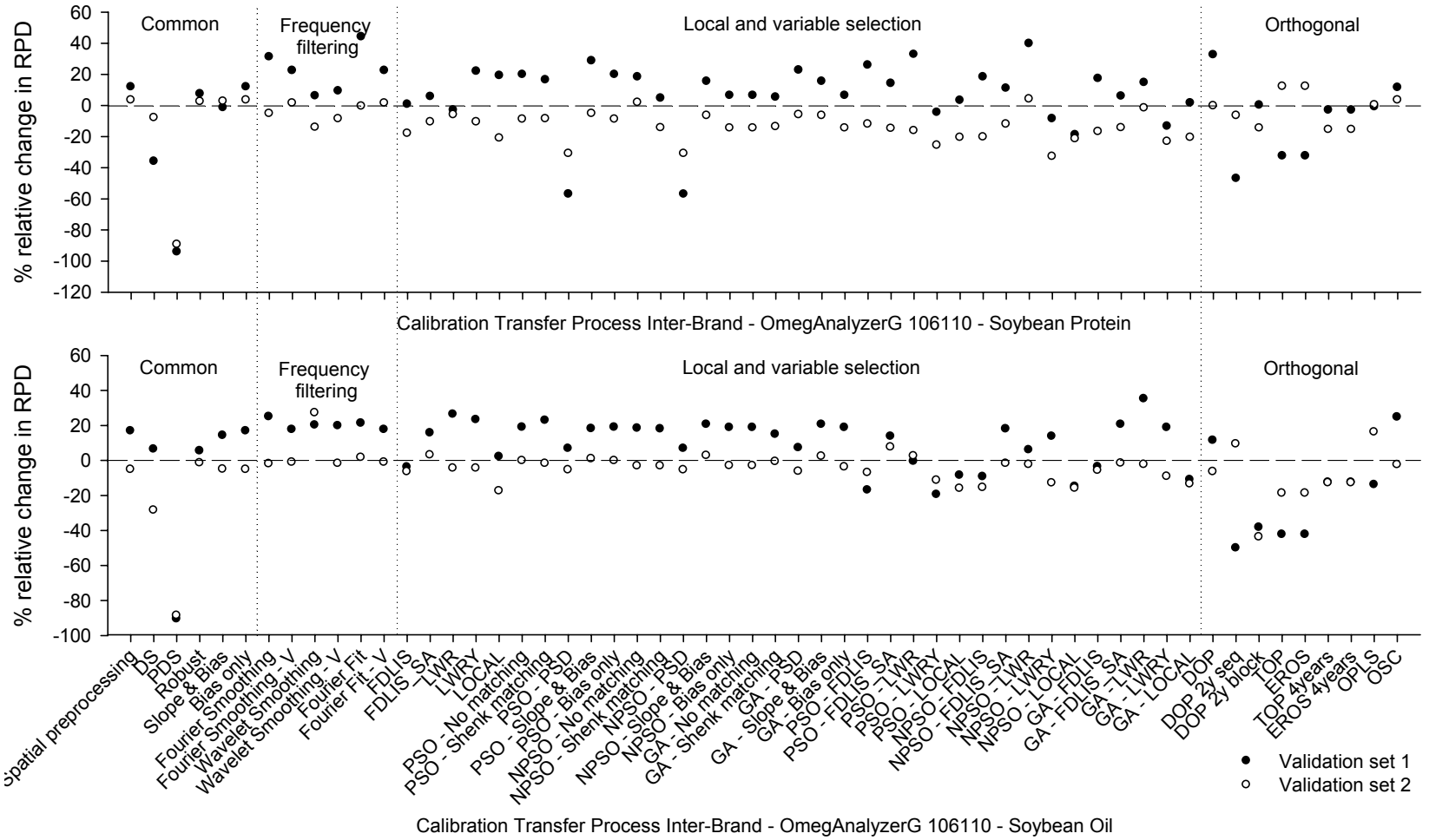


Figure 45: Comparison of inter-brand calibration transfer for OmegAnalyzerG 106110 (Infratec 1241 master) with original calibration results. V stands for when validation sets were preprocessed.

Figures 46 to 49 present the comparison between network masters and transferred models in intra-brand (figures 46 and 47) and in inter-brand (figures 48 and 49) calibration transfer scenarios.

In intra-brand situations, protein appeared to be easier to transfer than oil for infratec 1229 while it was the contrary for OmegAnalyzerG 106110. A larger difference between validation sets was observable for OmegAnalyzerG 106110 which might mean that while instruments appeared closer than Infratecs, a possible environmental effect impacted standardization trials. Large differences existed between master and secondary units with sample and variable selection techniques. On the contrary to some orthogonal methods that gave poor transferred performances, local and variable selection based methods differences arose from a large improvement of the calibration models on the master units and an obvious over-fitting that limited the transferability of the models. Many RPD were larger than 15.

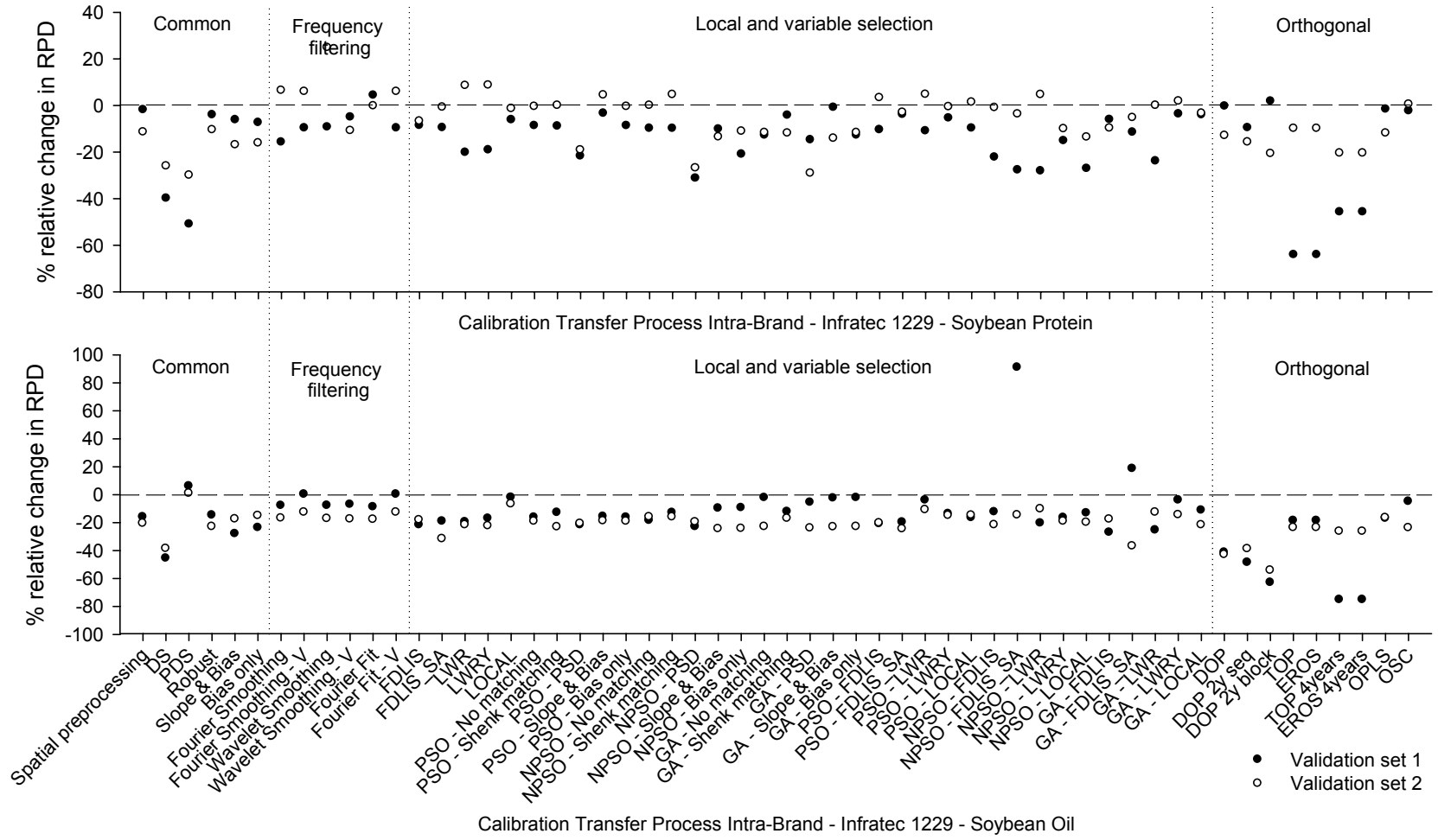


Figure 46: Comparison of intra-brand calibration transfer for Infratec 1229 with network master performances (Infratec 1241 master). V stands for when validation sets were preprocessed.

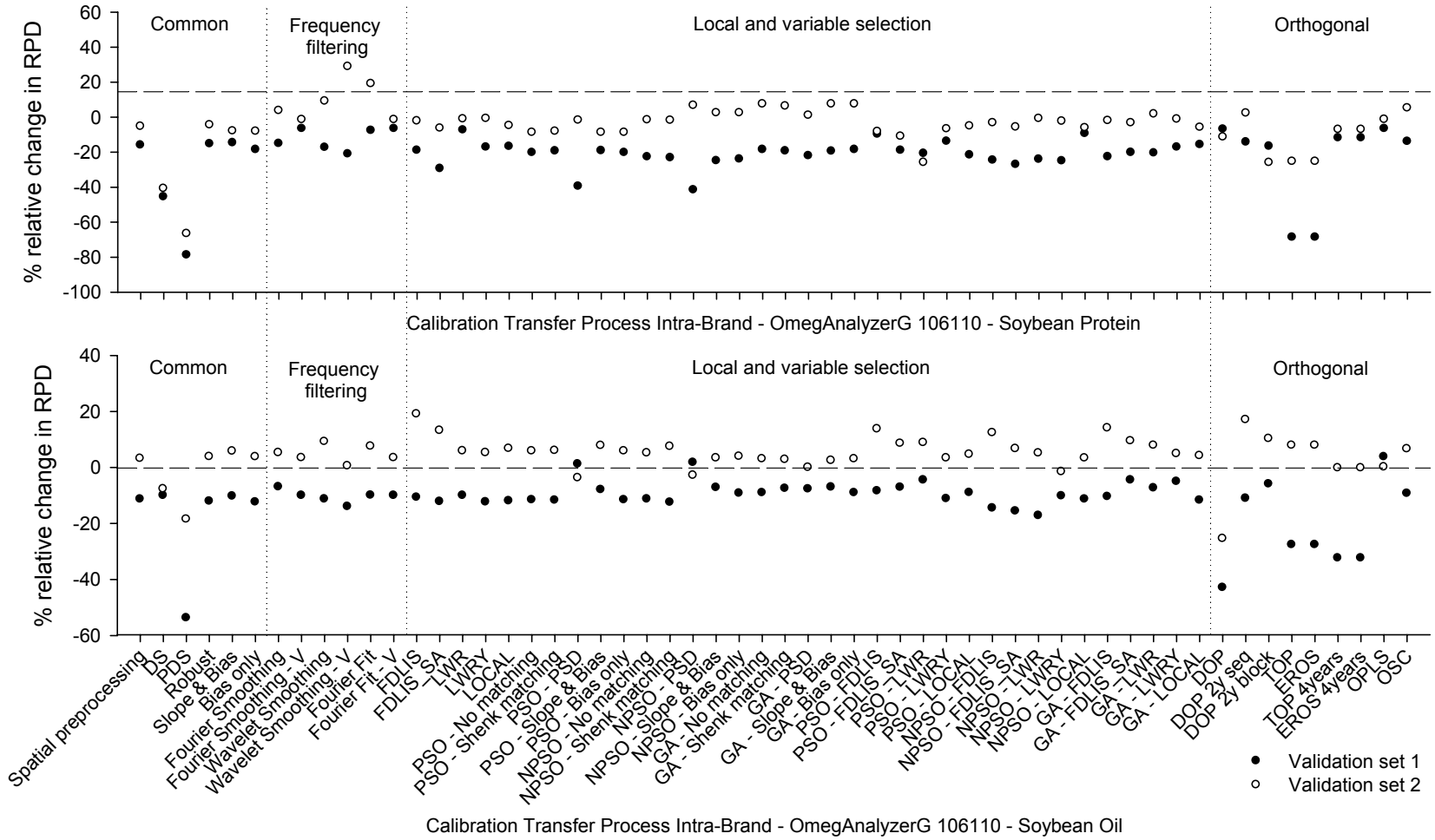


Figure 47: Comparison of intra-brand calibration transfer for OmegAnalyzerG 106110 with network master performances (OmegAnalyzerG 106118 master). V stands for when validation sets were preprocessed.

In inter-brand situations, very similar observations could be made with intra-brand calibration transfers. Local and variable selection methods suffered from over-fitting and their transfer abilities were limited.

Among all, local methods appeared to be the most unstable for calibration transfer. Even though they gave the best results in calibration development, they specialized calibration sets and limited the ability to perform successfully in external validation situations. The inherent risk they carry is too large for calibration transfer situations. Other methods such as spatial and frequency based filtering methods and orthogonal techniques, should be considered when designing intra and inter calibration transfer processes.

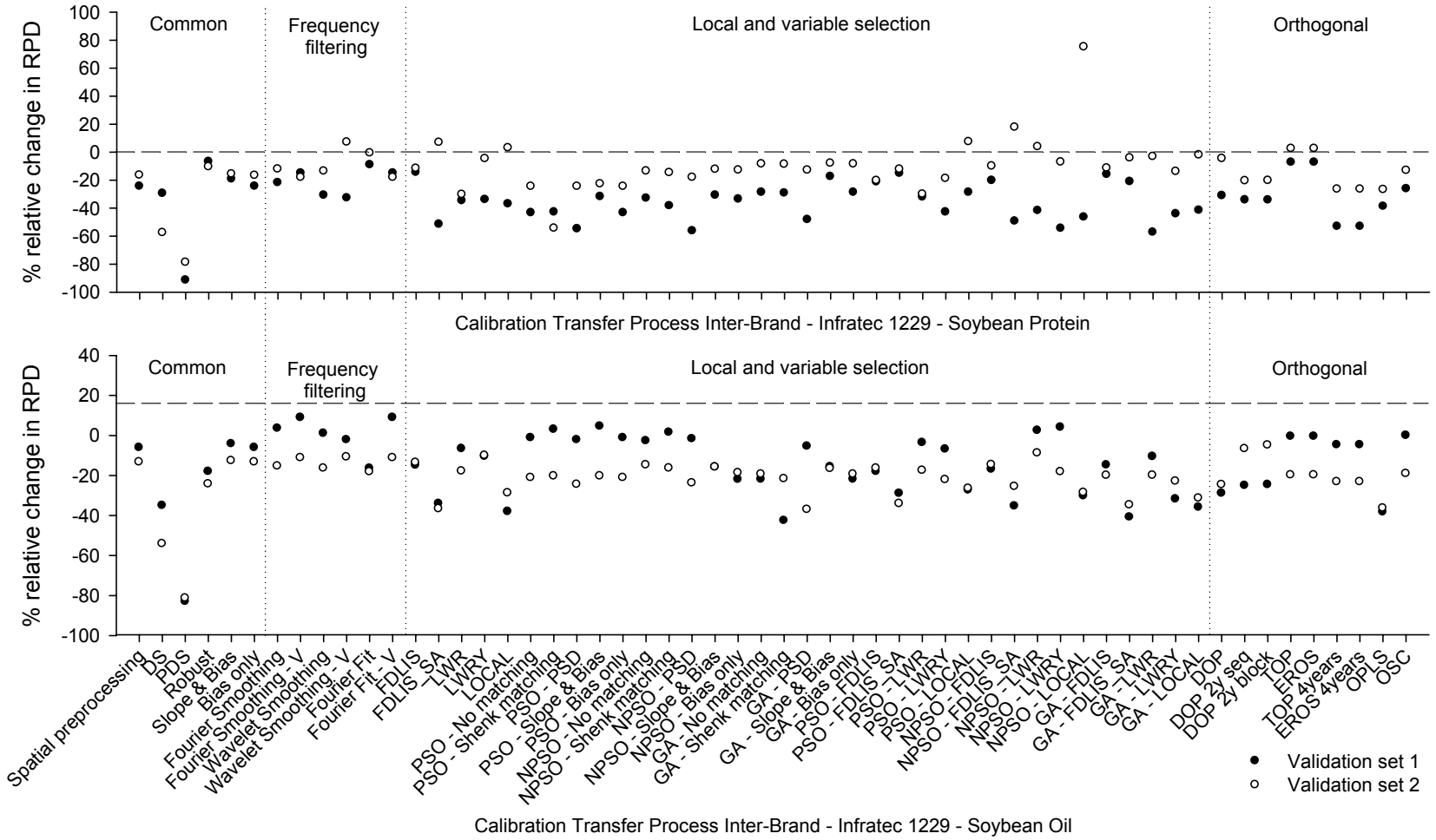


Figure 48: Comparison of inter-brand calibration transfer for Infratec 1229 with network master performances (Infratec 1241 master). V stands for when validation sets were preprocessed.

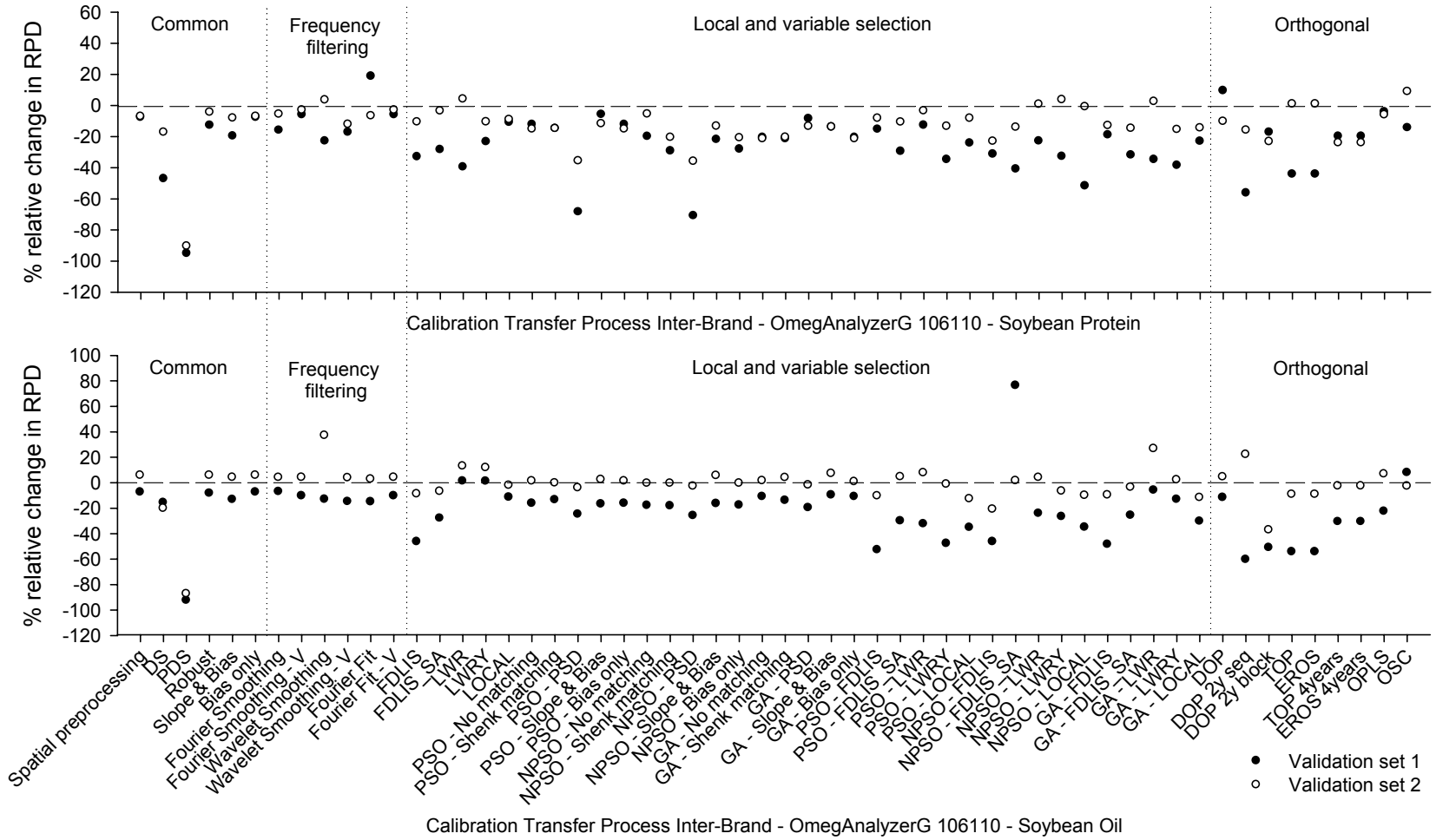


Figure 49: Comparison of inter-brand calibration transfer for OmegAnalyzerG 106110 with network master performances (OmegAnalyzerG 106118 master). V stands for when validation sets were preprocessed.

CHAPTER 6.

GENERAL CONCLUSION

In this study, we evaluated the possibility to filter, select samples and/or variables, and remove orthogonal signal for the calibration development and transfer of protein, oil, and linolenic acid models. The use and effectiveness of most methods was parameter, instrument, and validation set dependent. Large improvements were observed in calibration development where local methods provided the best performances. Models were transferable across units of the same brand and of different brands. However, the loss of precision was larger for local models and optical methods. Local models tent to over-fit the calibration set and reduced the capacity of the models to handle a different variability than the one present in the calibration set.

The variety of techniques presented, developed, and applied in this dissertation provides the literature with a comparison of calibration and standardization in intra and inter-brand situations. It also creates the possibility for public and private organizations to evaluate the inclusion of new instruments in their historical. Compiled in a toolbox, these tools could be used to evaluate the possibility to transfer a calibration model on different instrument brands and would also provide chemometricians with knowledge about the methods they should avoid with specific types of instruments.

However, the complexity of some of the methods used (Fourier and wavelet transform, some spatial pretreatments) restrains their direct use. For instance, none of the four units considered in this study could operate local methods, and all the methods involving the modification of the samples to predict (orthogonal methods based on y, advanced spatial preprocessing). They would require custom made software to extract spectra and predict in real time. Instrument manufacturers will need to provide chemometricians with tools that can handle advanced calculations to be efficiently used in demanding applications.

REFERENCES

Peer reviewed articles and books

Abdi H., 2003. *Least-squares. In* Encyclopedia for research methods for the social sciences, Lewis-Beck M., Bryman A., Futing T., (editors), Thousand Oaks, CA, USA.

Alsberg B.K., Woodward A.M., Winson M.K., Rowland J., Kell D. B., 1997. Wavelet Denoising of Infrared Spectra, *Analyst*, 122:645–652.

Andrew A., Fearn T., 2004. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation, *Chemometrics and Intelligent Laboratory Systems*, 72:51-56.

Andrews D.T., Wentzell P.D., 1997. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer, *Analytica Chemica Acta*, 350:341-352.

AACC International, 2000. Approved Methods of the American Association of Cereal Chemists, 10th Ed. Methods 39-00. The Association: St. Paul, MN.

AOAC Method 990.03. Official Methods of Analysis, 1990. Association of Official Analytical Chemists, 15th Edition

AOCS Method Ac 3-41. Approved Methods of the American Oil Chemist's Society, 1997.

Barnes R.J., Dhanoa M.S., Lister S.J., 1989. Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, *Applied Spectroscopy*, 43(5):772-776.

Berardo N., Pisacane V., Battilani P., Scandolara A., Pietri A., Marocco A., 2005. Rapid detection of kernel rots and mycotoxins in maize by near-infrared reflectance spectroscopy, *Journal of agricultural and food chemistry*, 53(21):8128-8134.

Bokobza L., 2002. Origin of Near-Infrared Absorption Brands. *In* Near-Infrared Spectroscopy: Principles, Instruments, Applications, Siesler H.W., Ozaki Y., Kawata S., Heise H.M. (editors), WILEY-VCH Verlag GmbH, Weinheim, Germany.

Bouveresse E. Massart D.L., Dardenne P., 1994. Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardization samples, *Analytica Chimica Acta*, 297:405-416.

Bouveresse E., Hartmann C., Marrart L., 1996. Standardization of Near-Infrared Spectrometric Instruments, *Analytical Chemistry*, 68:982-990.

Bouveresse E. Massart D.L., 1996. Standardization of near-infrared spectrometric instruments: A review, *Vibrational Spectroscopy*, 11:3-15.

Boysworth M.K., Booksk K.S., 2001. Aspects of Multivariate Calibration Applied to Near-Infrared Spectroscopy. *In* Handbook of Near-Infrared Analysis, Burns D. A. and Ciurczak E. W. (editors), 2nd ed., Marcel Dekkker, Inc., New York, NY, USA.

Brumm T., 2007. Personal communication.

Burns D.A., Ciurczak E.W. (editors), 2008. Handbook of Near-Infrared Analysis, 3rd ed., Taylor and Francis, Boca Raton, Florida, USA.

Chang S.Y., Baughman E.H., McIntosh B.C., 2001. Implementation of Locally Weighter Regression to Maintain Calibrations of FT-NIT Analyzers for Industrial Processes, *Applied Spectroscopy*, 55:1199-1206.

Chauchard F., Cogdill R. Roussel S., Roger J.M., Bellon-Maurel V., 2004. Application of LS-SVM to non-linear phenomena in NIR spectroscopy : development of a robust and portable sensor for acidity prediction in grapes, *Chemometrics and Intelligent Laboratory Systems*, 71:141-150.

Ciurczak E.W., 2001. Principle of Near-infrared Spectroscopy, *In* Handbook of Near-Infrared Analysis, Burns D.A. Ciurczak E.W. (editors), 2nd ed., Marcel Dekker, Inc, New York, USA.

Cleveland W.S., Devlin S.J., 1988. Locally Weighted Regression: an Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, 83:596-610.

Cogdill R.P., Hurburgh C.R., 2002. Local Chemometrics, 11th International Diffuse Reflectance Conference, Chambersburg, PA, USA.

Davis A.M.C., 1987. Genetic Algorithms and Simulated Annealing, Pitman, London, UK.

Davies A.M.C., Britcher H.V., Franklin J.G., Ring S.M., Grant A., McClure W.F., 1988. The Application of Fourier Transformed NIR Spectra to Quantitative Analysis by Comparison of Similar Indices (CARNAC), *Mikrochimica Acta* (Wien) 94:61-64

Davies A.M.C., Fearn T., 2006. Quantitative analysis via near infrared databases: comparison analysis using restructured near infrared and constituent data-deux (CARNAC-D), *Journal of Near Infrared Spectroscopy*, 14:403-411.

Dean T., Isaksson T., 1993. Standardization: What is it and how is it done? Part 2, *NIR News*, 4(2):14-15.

Despagne F., Walczak B., Massart D.L., 1998. Transfer of calibrations of near-infrared spectra by neural networks, *Applied Spectroscopy*, 52:732-745.

Duponchel L., Ruckebusch C., Huvenne J.P., Legrand P., 1999. Standardization of near-infrared spectrometers using artificial neural networks, *Journal of Near Infrared Spectroscopy*, 7:155-166.

Dou Y., Sun Y., Ren Y., Ren Y., 2005. Artificial neural network for simultaneous determination of two components of compound paracetamol and diphenhydramine hydrochloride powder on NIR spectroscopy, *Analytica Chimica Acta*, 528(1):55-61.

Esbensen K. H., 2002. Multivariate Data Analysis –in practice. An introduction to multivariate data analysis and experimental design, 5th edition, CAMO ProcessAS, Oslo, Norway.

Fearn T., 1996. Comparing Standard deviations". *N-I-R news*. **7**:5.

Fearn T., 2001. Local or global?, *NIR news*, 12(3):10-11.

Fearn T., 2001. Standardization and calibration transfer for near infrared instruments: a review, *Journal of near infrared spectroscopy*, 9:229-244.

Fearn T., Davies A.M.C., 2003. A comparison of Fourier and wavelet transform in the processing of near infrated spectroscopic data: Part 1. Data compression, *Journal of Near Infrared Spectroscopy*, 11:3-15.

Fernandez J., 2006. Personal communication.

Fernandez-Cabanas V.M., Garrido-Varo A., Perez-Martin D., Dardenne P., 2006. Evaluation of Pretreatmwent Strategies for Near-Infrared Spectroscopy Calibration Development of Unground and Ground Compound Feedingstuffs, *Applied Spectroscopy*, 60(1):17-23.

Feudale R.N., Woody N.A., Tan H., Myles A.J., Brown S.D., Ferre J., 2002. Transfer of multivariate calibration models: a review, *Chemometrics and Intelligent Laboratory Systems,* 64:181-192.

Fox G.P., Onley-Watson K., Osman A., 2002. Multiple Linear Regression Calibrations for Barley and Malt Protein Based on the Spectra of Hordein, *J. Inst. Brew.*, 108(2): 155–159.

Geladi, P. and Kowalski B.R., 1986. PLS Tutorial, *Anal. Chim. Acta.*, 185(1): 1-17.

Hammond E.G., 1991. Organization of Rapid Analysis of Lipids in Many Individual Plants. *In* Modern Methods of Plant Analysis, New Series, Volume 12, Essential Oils and Waxes, Springer-Verlag, p. 321.

Heckman R.A., Diffee J.T., Milhous L.A., 1987. Transfer of near-infrared monochromator calibrations for tobacco constituents to tilting-filter instruments, *Analytica Chimica Acta*, 192:197-203.

Hopking D.W., 2008. Personal communication.

Hurburgh C.R., Brumm T., 2007. Grain Quality, in Managing Grain After Harvest, course material, Iowa State University.

Igne B., Gibson L.R., Rippke G.R., Hurburgh C.R., 2006. Evaluation of preprocessing methods in the development of near-infrared models for triticale protein and moisture, presented to the World Grains Summit: Foods and Beverages, San Francisco, CA, September 2006.

Igne B., Gibson L.R., Rippke G.R., Schwarte A., Hurburgh C.R., 2007a. Triticale Moisture and Protein Prediction by Near Infrared Spectroscopy, *Cereal Chemistry*, 84(4):328-330.

Igne B., Gibson L.R., Rippke G.R., Hurburgh C.R., 2007b. Influence of the yearly variability of agricultural products on the calibration process: a triticale example, *Cereal Chemistry*, 84(6):576-581.

Igne B., Rippke G.R., Hurburgh C.R., 2008. Measurement of Whole Soybean Fatty Acids by Near Infrared Spectroscopy, *Journal of the American Oil Chemist's Society*, 85:1105-1113.

Ingle J.D. Jr., Crouch S.R., 1988. Spectrochemical Analysis, Prentice-Hall, Inc., New Jersey, USA.

Katawa S., 2002. Instrumentation for Near-Infrared Spectroscopy. *In* Near-Infrared Spectroscopy: Principles, Instruments, Applications, Siesler, H. W., Ozaki Y., Kawata S., Heise H.M. (editors), WILEY-VCH Verlag GmbH, Weinheim, Germany.

Keckman V., 2001. Learning and Soft Computing, The MIT press.

Kennard R.W., Stone L.A. Computer aided design of experiments, *Technometrics*, 11:137-148.

Kennedy J., Eberhart R., 1995. Particle swarm optimization In: Proceedings of IEEE International Conference On Neural Networks, Perth, Australia, 1942–1948.

Kennedy J., Eberhart R.C., 2001. Swarm Intelligence, Academic Press, San Diego, CA, USA.

Kim S., Park J., Chun J.H. and Lee S.M., 2004. Determination of Ergosterol Used as a fungal Marker of Red Pepper (Capsicum annuum, L.) Powders by Near-Infrared Spectroscopy, *Food Science and Biotechnology*, 12 (3):257-261.

Kovalenko I., 2006. Near-Infrared (NIR) Spectroscopy and Chemometrics Primer, Grain Quality Laboratory internal document, Iowa State University.

Kovalenko I.V., Rippke G.R.,. Hurburgh C.R., 2006. Measurement of Soybean Fatty Acids by Near-Infrared Spectroscopy: Linear and Nonlinear Calibration Methods. *Journal of the American Oil Chemist's Society*, 83:421-427.

Leardi R., Boggia R., Terrile M., 1992. Genetic Algorithms as a Strategy for Feature Selection, *Journal of Chemometrics*, 6:267-282.

Levasseur C., 2004. La Spectroscopie proche infrarouge : un outil de détection du risque mycotoxique sur blé, Mémoire d'Ingénieur de l'ESA Purpan, Toulouse, France.

Lorber A., Wangen L. E. and Kowalski B. R., 1987. A Theoretical Foundation for the PLS Algorithm, *Journal of Chemometrics*, 1:19-31.

Lu J.X., Shen Q., Jiang J.H., Shen G.L., Yu R.Q., 2004. QSAR analysis of cyvlooxygenase inhibitor using particlw swarm optimization and multiple linear regression, *Journal of Pharmaceutical and Biomedical* Analysis, 35:679-687.

Luke, B.T., 1994. Evolutionary programming applied to the development of quantitative structure-activity relationships and quantitative structure-property relationships. *Journal of Chemical Information and Modeling,* 34, 1279–1287.

Marcott C., 2008. Personal communication.

Mallat S., 1999. A Wavelet tour of signal processing, Second edition, Academic Press.

Mark H., Workman J., 1988. A new Approach to generating transferable calibrations for quantitative near-infrared spectroscopy, *Spectroscopy*, 3(11):28-36.

Martens H. and Næs T., 1989. Multivariate Calibration, John Wiley & Sons, New York, NY, USA.

McClure W.F., 2001. Near-Infrared Instrumentation. *In* Near-Infrared Technology in the Agricultural and Food Industries, Williams P. and Norris K. (editors), 2nd ed., AACC Inc., St. Paul, MN, USA.

McClure W.F., 2008. Analysis Using Fourier Transforms. Handbook of Near-Infrared Analysis, Edited by D. A. Burns and E. W. Ciurczak, Third edition.

Miller C.E., 2001. Chemical Principles of Near-Infrared Technology. *In* Near-Infrared Technology in the Agricultural and Food Industries, Williams P. and Norris K. (editors), 2nd ed., AACC Inc., St. Paul, MN, USA.

Næs T., Isaksson T., Kowalski B., 1990. Locally Weighted Regression and Scatter Correction for Near-Infrared Reflectance Data, *Analytical Chemistry*, 62(7):664 - 673.

Næs T., Isaksson T., Fearn T., Davies T., 2002. A user-friendly guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK.

Norgaard L.; Saudland A.; Wagner J.; Nielsen J.P.; Munck L.; Engelsen S.B., 2000. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy, *Applied Spectroscopy*, 54(3):413-419

Paynter L.N., Hurburgh C.R., 1983. Reference methods for corn moisture determination, American Society of Agricultural Engineers, 83:37-43.

Park B., Chen Y.R., Hruschka W.R., Shackelford S.D., Koohmaraie M., 2001. Principal Component regression of Near-Infrared Reflectance spectra for Beef Tenderness Prediction, *Transactions of the American Society of Agricultural Engineers*, 43(3):609-615.

Peinado A.C., van den Berg F., Blanco M., Bro R., 2006. Temperature-induced variation for NIR tensor-based calibration, *Chemometrics and Intelligent Laboratory Systems*, 83:75–82.

Pitman, E.J.G., 1939. A note on normal correlation, *Biometrika*, 31:9-12.

Roger J.M., Chauchard F., Bellon-Maurel V., 2003. EPO-PLS external parameter orthogonalization of PLS. Application to temperature-independent measurement of sugar content of intact fruits, *Chemometrics and Intelligent Laboratory Systems*, 66:191-204.

Savitzky A. and Golay M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Analytical Chemistry*, 36:1627–1639.

Shen Q., Jiang J.H., Jiao C.X., Shen G.L., Yu R.Q., 2004. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists, *European Journal of Pharmaceutical Sciences*, 22:145-152.

Shenk J.S., Westerhaus M.O., Templeton W.C., 1985. Calibration Transfer Between Near Infrared Reflectance Spectrometers, *Crop Science*, 25:159-161.

Shenk J.S., Westerhaus M.O., 1993. Comments on Standardization: Part 2, *NIR news*, 4(5):13-15.

Shenk J.S., Westerhaus M.O., Berzaghi P., 1997. Investigation of a LOCAL calibration procedure for near infrared instruments, *Journal of Near Infrared Spectroscopy*, 5:223-232.

Shi B., Ji B., Zhu D., Tu Z., Qing Z., 2008. Study on genetic algorithms-based NIR wavelength selection for determination of soluble solids content in fuji apples, *Journal of Food Quality*, 31(2):232-249.

Siesler H.W., Ozaki Y., Kawata S., Heise H.M. (editors), 2002. Near-Infrared Spectroscopy: Principles, Instruments, Applications, WILEY-VCH Verlag GmbH, Weinheim, Germany.

Siska J., Hurburgh C.R., and Siska P., 2001. The standardization of near-infrared instruments using master selection and Wiener filter methods. *Journal of Near Infrared Spectroscopy*, 9:97-105.

Sun J., 1996. A Multivariate Principal Component Regression Analysis of NIR Data, *Journal of Chemometrics*, 10:1-9.

Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., Vandewalle J., Least Squares Support Vector Machines, World Scientific, Singapore, 2002.

Swierenga H., de Groot P.J., de Weijer A.P., Derksen M.W.J., Buydens L.M.C., 1998. Improvement of PLS model transferability by robust wavelength selection, *Chemometrics and Intelligent Laboratory Systems*, 41:237-248.

Tan H.W., Brown S.D., 2001. Wavelet hybrid direct standardization of near-infrared multivariate calibrations, *Journal of Chemometrics*, 15:647-663.

Teppola P., Mujunen S.P., Minkkinen P., 1999. Kalman filter for updating the coefficients of regression models. A case study from an activated sludge waste-water treatment plant, *Chemometrics and Intelligent Laboratory Systems*, 45:341-384.

Trygg J., Wold S., 2002. Orthogonal projections to latent structures (O-PLS), *Journal of Chemometrics*, 16:119-128.

Vapnik V., 1995. The nature of Statistical Learning Theory, Springer Verlag, New York.

Wang Y., Kowalski R., 1993. Temperature-compensating calibration transfer for near-infrared filter instruments, *Analytical Chemistry*, 65:1301-1303.

Wang Y., Veltkamp D.J., Kowalski R., 1991. Multivariate Instrument Standardization, *Analytical Chemistry*, 63:2750-2756.

Walmsley A.D., 2006. Statistical Evaluation of Date. In Practical Guide to Chemometrics, Gemperline P. (Editor), 2nd ed., CSR Press, Boca Raton, FL, USA.

Williams P. and Norris K. (editors), 2001. Near-Infrared Technology in the Agricultural and Food Industries, 2nd ed., AACC Inc., St. Paul, MN, USA.

Williams P., 2001. Implementation of Near-Infrared Technology. *In* Near-Infrared Technology in the Agricultural and Food Industries, Williams P. and Norris K. (editors), 2nd ed., AACC Inc., St. Paul, MN, USA.

Wise B.M., Gallagher N.B., Bro R., Shaver J.M., Windig W., Koch R.S., 2006. PLS_Toolbox Version 4.0 for use with MATLAB™, Eigenvector Research, Inc. Wenatchee, WA, USA.

Wold H., 1966. Estimation of principal components and related models by iterative least squares. *In* P.R. Multivariate Analysis, Krishnaiaah (editor), New York: Academic Press, USA.

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

Wold S., Ruhe A., Wold H., and Dunn W.J.III, 1984. The Collinearity Problem in Linear Regression, The Partial Least Square Approach to Generalized Inverses", *SIAM Journal of Scientific and Statistical Computing*, (5):735-743.

Workman J.J. Burns D.A., 2001. Commercial NIR Instrumentation. *In* Handbook of Near-Infrared Analysis, Burns D. A. and Ciurczak E. W. (editors), 2nd ed., Marcel Dekkker, Inc., New York, NY, USA.

Wu W., Walczak B., Massart D.L., Heuerding S., Erin F., Last I.R., Prebble K.A., 1996. Artificial neural networks in classification of NIR spectral data: Design of the training set, *Chemometrics and Intelligent Laboratory Systems,* 33(1):35-46.

Xie Y., Hopke P.K., Paatero P., 1999. Calibration transfer as a data reconstrunction problem, *Analytica Chemica Acta*, 384:193-205.

Yu L., Jiang J.H., Wu H.L., Shen G.L., Yu R.Q., 2007. Variable-weighted PLS, *Chemometrics and Intelligent Laboratory Systems*, 85:140-143.

Zeaiter M., Roger J.M., Bellon-Maurel V., 2006. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations, *Chemometrics and Intelligent Laboratory Systems*, 80:227-235.

Zhu Y., Fearn T., Samuel D., Dhar A., Hameed O., Bown S.G., and Lovat L.B., 2008. Error removal by orthogonal subtraction (EROS): a customised pre-treatment for spectroscopic data, *Journal of Chemometrics*, 22:130-134.

Internet references:

Davis A.M.C., 2007. Thread: Can PDS give worse results than no standardization?
http://www. nirpublications.co.uk/cgi-bin/discus/discus.cgi
visited 06/03/2008.

Jackson M., 2001. Fourier Transform
http://www.see.ed.ac.uk/~mjj/dspDemos/EE4/tutFT.html
06/04/2008.

Food and Drug Administration, 2007.
http://www.fda.gov/consumer/ updates/melamine051407.html
06/03/2007.

Food and Drug Administration, 2007.
http://www.fda.gov/oc/opacom/hottopics/spinach.html
06/03/2007.

The USA PATRIOT Act, 2005.
http://www.epic.org/privacy/terrorism/usapatriot/
06/03/2007.

United State Department of Agriculture, 2007. World Agricultural Production
http://www.fas.usda.gov/psdonline/psdHome.aspx
08/26/2008.