Multivariate NIR Studies of Seed-Water Interaction in Scots Pine Seeds (*Pinus sylvestris L.*)

Torbjörn Lestander
Department of Silviculture
Umeå

Doctoral Thesis
Swedish University of Agricultural Sciences
Umeå 2003

Acta Universitatis Agriculturae Sueciae
Sylvestria 282
ISSN: 1401-6230
ISBN 91-576-6516-8

Printed by: Grafiska Enheten, SLU, Umeå, Sweden. 2003.

Abstract

Lestander, T. A. 2003. Multivariate NIR studies of seed-water interaction in Scots pine seeds (*Pinus sylvestris* L.). Doctoral dissertation.
ISBN 91-576-6516-8, ISSN: 1401-6230

This thesis describes seed-water interaction using near infrared (NIR) spectroscopy, multivariate regression models and Scots pine seeds. The presented research covers classification of seed viability, prediction of seed moisture content, selection of NIR wavelengths and interpretation of seed-water interaction modelled and analysed by principal component analysis, ordinary least squares (OLS), partial least squares (PLS), bi-orthogonal least squares (BPLS) and genetic algorithms.

The potential of using multivariate NIR calibration models for seed classification was demonstrated using filled viable and non-viable seeds that could be separated with an accuracy of 98-99%. It was also shown that multivariate NIR calibration models gave low errors (0.7% and 1.9%) in prediction of seed moisture content for bulk seed and single seeds, respectively, using either NIR reflectance or transmittance spectroscopy. Genetic algorithms selected three to eight wavelength bands in the NIR region and these narrow bands gave about the same prediction of seed moisture content (0.6% and 1.7%) as using the whole NIR interval in the PLS regression models. The selected regions were simulated as NIR filters in OLS regression resulting in predictions of the same quality (0.7 % and 2.1%). This finding opens possibilities to apply NIR sensors in fast and simple spectrometers for the determination of seed moisture content.

Near infrared (NIR) radiation interacts with overtones of vibrating bonds in polar molecules. The resulting spectra contain chemical and physical information. This offers good possibilities to measure seed-water interactions, but also to interpret processes within seeds. It is shown that seed-water interaction involves both transitions and changes mainly in covalent bonds of O-H, C-H, C=O and N-H emanating from ongoing physiological processes like seed respiration and protein metabolism. I propose that BPLS analysis that has orthonormal loadings and orthogonal scores giving the same predictions as using conventional PLS regression, should be used as a standard to harmonise the interpretation of NIR spectra.

Key words: Single seed, near infrared spectroscopy, reflectance, transmittance, multivariate analysis, wavelength selection, PCA, OLS, PLS, bi-orthogonal PLS, interval PLS, genetic algorithms, seed viability, seed moisture content.

Author's address: Torbjörn Lestander, Department of Silviculture, Swedish University of Agricultural Sciences, SE - 901 83 Umeå, Sweden. E-mail: torbjorn.lestander@omv.slu.se

to

Ylva , Ragna, Hedvig

Contents

Appendix

List of original papers

This thesis is based on the following papers, which will be referred to in the text by their respective Roman numerals.

I Lestander, T.A. and Odén, P.C. 2002. Separation of viable and non-viable filled Scots pine seeds by differentiating between drying rates using single seed near infrared transmittance spectroscopy. *Seed Science and Technology*, **30** (2): 383-392.

II Lestander, T.A. and Geladi, P. 2003. NIR spectroscopic measurement of moisture content in Scots pine seeds. *Analyst*, **128** (4): 389-396.

III Lestander, T.A., Leardi, R. and Geladi, P. 2003. Selection of NIR wavelengths by genetic algorithms for the determination of seed moisture content. (submitted).

IV Lestander, T.A., Geladi, P. 2003. How does multivariate regression predict moisture content from NIR spectra of seeds? (submitted).

Introduction

Seeds are fundamental for regeneration of most plant species. In agriculture and forestry seeds are the starting point for the production of crops for food, feed and raw industrial materials. Seeds are also directly or indirectly the base for the lives of 99% of the world's human population (Urmstrom 1997). Globally, in agriculture about 664 million hectares of cereals alone are harvested annually (FAOSTAT 2003). The main species are wheat, rice and maize followed by barley, sorghum, millet, oats and rye. In forestry more than 3 million hectares are annually regenerated mainly by species from the genera *Pinus* and *Eucalyptus* (FAO 2001).

To sow a seed is to expect that it will germinate and develop into a plant. This does not always happen due to many biotic and abiotic factors or perhaps due to the simple reason that the seed is non-viable (dead). Therefore it is of importance to get rid of dead seeds. Already in the times of the Old Testament, farmers used simple techniques like throwing seeds in the wind (Eskeröd 1973) to clean and sort seeds. Seeds are valuable resources and in order to use seeds as efficiently as possible, sophisticated and highly mechanised techniques have been developed to clean and sort seeds before sowing.

The most used techniques for screening of seeds are based on physical properties like width, thickness and length but also specific gravity, shape and colour (Harmond *et al*. 1968). Round holes are used to sort seeds according to their width. Rectangular shaped holes separate seeds of different thickness. Sorting based on length is more complicated. An often-used technique is to place seeds in a rotating horizontal cylinder that has round concave cavities on the inside. Seeds that are shorter than the diameter of a cavity will attach to it. When the cavity turns upside down during the rotation cycle the sorted seeds drop into a collection channel inside the cylinder and are removed, while larger seeds are retained. On so called gravity tables air is blown from below the seeds lying on an oscillating tilted plane. For each cycle those seeds that are in contact with the plane are pushed upwards when the oscillation reaches maximum in the vertical plane. Seeds with low specific gravity will float in air and not be pushed as often as seeds of high specific gravity and therefore successively move downwards on the tilted plane. Gravity sorting is also done in airstreams, liquids of different specific weights (e.g. Falleri & Parcella 1997) or in laminar water streams (Bergsten 1987, Lestander 1988). By rolling or gliding down a tilted plane seeds with different forms can be sorted (Harmond *et al*. 1968).

Colour sorters represent a different approach to seed sorting (Figure 1). By this technique single seeds are characterised by spectrometric means in the visible wavelength region (Anon. 2002a and 2002b). Single seeds glide through a tilted channel and at the end of the channel every seed is illuminated by a radiation source. The reflectance from the seed is detected and depending on the reflected radiation a decision algorithm decides whether an air ejector should blow a pulse of compressed air or not. If a short air pulse is blown the seed will be pushed aside and fall down in a channel for rejects. This technique has been commercially

available for nearly half a century (Powers *et al.* 1953). Besides reflectance of seeds, transmittance through seeds has also been used for sorting of almonds (Pearson 1999). A new variant of colour sorting is to use a laser beam to measure the amount of chlorophyll fluorescence in seeds mainly to remove immature green grains (Jalink *et al.* 1998, Konstantinova *et al.* 2002, Anon. 2002c).

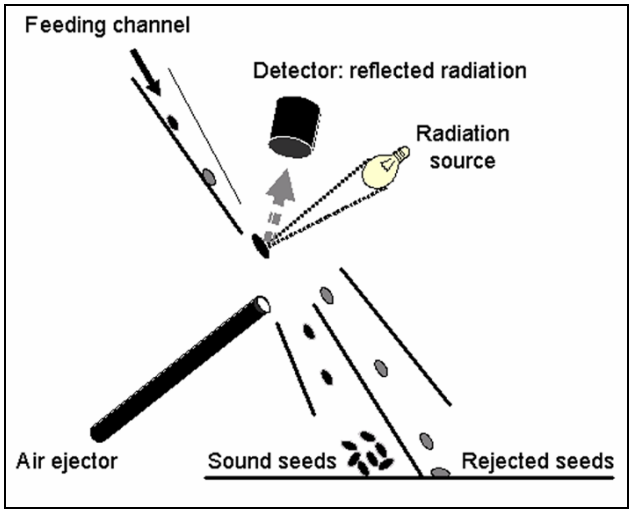

Figure 1. Principle for colour sorting of seeds.

Most often there is a low to moderate correlation between sorting criteria and biological target properties such as ability to germinate. Optimum viability can be obtained by accumulation of just germinated seeds (Hagner 1981) in combination with sowing of pre-germinated seed (Salter 1978). This has been tried, but with low success rate as perishable pre-germinates are more difficult to store and handle than intact seeds. Another approach is to develop artificial seeds by using somatic embryos (*e.g.* von Arnold *et al.* 2002). This is promising for some species, for example in mothbean (Malabadi & Nataraja 2002).

Low germination capacity of a seed lot can be compensated by sowing more seeds in order to obtain the same amount of plants. Compensation sowing in the field may result in uneven stand densities (Hühn 2001) that need thinning. This problem is still more accentuated when producing containerized plants. These are widely used for reforestation in northern Europe, in for example Nordic forestry. In Sweden alone about 300 million such plants of conifers are annually produced for planting (Hannerz *et al.* 2000). In containerized plant production the sowing of two seeds per container reduces the amount of "empty" containers, but results also in having two plants in many of the containers. For example if the germination capacity is 80-90% double seed sowing gives double plants in 64-81% of the containers. Thus, thinning within containers has to be carried out as double plants may give root deformations (Nyström 1982**).** This is costly in highly mechanised plant production systems, but empty containers also give high handling and transport costs. The reduction of empty containers is in this case, using double seed sowing, down from 10-20% to 1-4%. This gives 9-16% more plants at the price of thinning and sowing double the amount of seeds. A more cost effective

and sustainable way seems to be to increase the germination capacity by removal of seeds that do not germinate.

The ultimate goal is to achieve 100% germination of otherwise high quality seeds. This goal governs also a technical progress including precision sowing in the field **(**Kachman & Smith 1995, Bracy & Parish 2001, Ozmerzi *et al.* 2002**)** and single seed sowing. The question is how to reach 100% germination. Is it possible to separate two seeds that look identical, if one is dead and the other is alive? How can one find a direct relationship between fast measurements and biological properties of seeds? One way is to enhance biological processes in seeds by creating conditions that promote progress towards germination.

According to Bewley & Black (1994) germination begins with water uptake by the seed (imbibition) and ends with the start of elongation by the embryonic axis, usually the radicle. During germination and early seedling development the food reserves that are stored in the endosperm are mobilized by hydrolytic enzymes and transformed and transported by different processes, finally reaching the embryo or early seedling mainly as amide from stored proteins and as sucrose from stored carbohydrates or lipids (Copeland & McDonald 2001). Germination is an important step in the transition from an embryo dependent on stored reserves to a photosynthesizing autotrophic plant. Besides light, oxygen and temperature, water is a major factor that controls seed germination in many species and thus the biological processes in seeds. It has also been shown that imbibed viable and non-viable seeds have different drying rates (Simak 1984) and this phenomenon is used in large scale to increase germination capacity of conifer seeds (Lestander 1986, 1988). Thus, water shows a direct relationship with biological properties of seeds and additional questions are therefore: What is the role of water in these relations? How can one measure seed-water interaction in a flow of single seeds without damaging the seeds? These questions are the basis of this thesis. A major goal in studying these problems is to produce fundamental and applied knowledge of seed-water interactions useful in practical applications.

The role of water in seed management

The seed-water interaction is closely linked to temperature and time as described by the concept of hydrothermal time. In the concept of degree-days, also called thermal time, only seed germination time courses at a given water potential and suboptimal temperatures are described. The water potential is often presupposed as free access to water. The hydrotime model describes only the effect of water potential on germination time at a given temperature. One example of this is a standard germination test where seeds at a given temperature absorb water via a germination paper that is connected to a given water level (ISTA 1999).

The hydrothermal time model describes seed germination at different combinations of reduced water potentials and sub/supra-optimal to optimal temperatures. This concept, first used by Gummerson (1986), has been further developed by Finch-Savage and coworkers (1998, 2000). The unit for

hydrothermal time is water potential (**Ψ**, unit: MPa) times degrees centigrade (**T**, unit: ºC) times time (MPa ºC d, where d is days).

The advantage of hydrothermal time is that this concept links water, temperature and time and defines the limits for three different seed states due to the combination of water and temperature. In the first state (Q) the seeds become quiescent and can not progress towards germination, in the second state (P) seeds can progress towards germination but radicle emergence can not occur, and finally, in the third state (R) seeds can progress to radicle emergence and germinate (Figure 2).

These seed states correspond to the observed three phases of water uptake in seeds described in the seed literature *e.g* Bewley & Black (1994) and Kigel & Galili (1995). In Phase 1 water is transported into or out of seeds without any major biological activity occurring in the seed. In this phase the seeds act like a sponge or a piece of wood. When the water content is high enough the seeds enter Phase 2 and cascades of biological processes start. Seed respiration increases and proteins and DNA are synthesised – the seed becomes a biological active organism. One may say that Phase 2 initiates the transition of an embryo dependent on stored nutrients into a plant that is autotrophic by photosynthesis. Phase 3 starts when the radicle emerges mainly due to cell elongation caused by increased water uptake. At the time of radicle protrusion the moisture content in seed embryos of maize increased to 55% of fresh embryo weight whereas the whole seed moisture content was ca 31-34% (McDonald *et al*. 1994). An intensive cell division starts in the embryo of the germinated seed, mainly in the meristem of the root tip, but before radicle emergence these cells are arrested in the cell cycle (de Castro *et al*. 2000) before cytokinesis (cell division). In conifer seeds however, cell division may precede radicle emergence (Bewley & Black 1994).

For seeds in Phase 2 of water uptake the seed respiration rate can be divided into two stages (Bewley & Black 1994). In the first, mitochondrial enzymes involved in the tricarboxylic acid (TCA) cycle are activated and respiration increases linearly with hydration of seed tissues followed by an increase in O_2 absorption. In the second stage, newly synthesized mitochondria and enzymes become limiting factors and the increase in O_2 absorption slows down. When seeds enter Phase 3, *i.e.* the seed germinate and cells divide, a second fast increase in O_2 absorption occurs. Respiration, *i.e.* the absorption of oxygen, is a part of the catabolism that is divided into three stages (Alberts *et al.* 1994): (i) breakdown of macromolecules into simple subunits; (ii) breakdown of subunits to acetyl-CoA and production of limited amounts NADH and ATP; (iii) complete oxidation of acetyl-CoA and production of large amounts of NADH via the TCA cycle and ATP via oxidative phosphorylation. ATP is an important energy source for biosynthesis.

Figure 2 presents the principle of the hydrothermal concept. At temperatures lower than $\mathbf{T}_{min}$ or higher than $\mathbf{T}_{max}$ the seeds become quiescent, *i.e.* none of the germination processes is taking place (Bewley & Black 1994). This happens also at water potential lower than $\mathbf{\Psi}_{min}$. This phenomenon is used in long time storage at low temperatures of both orthodox and recalcitrant seeds *i.e.* seeds that survive severe drying and seeds that are damaged by moderate drying, respectively. At

higher temperatures (but still lower than $\mathbf{T}_{max}$) and higher water potentials seeds progress towards germination.

Radicle emergence can be prevented if the water potential is lower than the base water potential for germination ($\mathbf{\Psi}_b$) or if the temperature is either lower than base temperature for germination ($\mathbf{T}_b$) or higher than the ceiling temperature ($\mathbf{T}_c$) for germination (Figure 2). The phenomenon of inhibition of radicle emergence at these seed states is widely used in many different methods to prime seeds before sowing (reviewed by Taylor *et al.* 1998). The main advantage of seed priming is that the germinating processes in the primed seeds will reach about the same level resulting in fast and even germination when sown in favourable germination conditions (Bray 1995). The seed water status can be regulated to suitable levels below base water potential for germination ($\mathbf{\Psi}_b$) using controlled hydration (Thomas 1983), solutions at specified water potential (Heydecker 1975) or solid matrix priming (*e.g.* Wu *et al.* 2001) consisting of solid particulate systems. Temperatures below $\mathbf{T}_b$ in wet conditions are also used as a seed priming method for many species and cold-wet treatments are mainly used to break seed dormancy (*e.g.* Downie *et al.* 1998). High temperature treatments are in some cases needed to break dormancy in species that deposit seeds in the soil and that germinate after forest fires (*e.g.* Granström & Schimmel 1993). There are also priming methods that allow favourable temperature and water regimes for radicle emergence ($\mathbf{T}_c > \mathbf{T} > \mathbf{T}_b$ and $\mathbf{\Psi} > \mathbf{\Psi}_b$), but then the treatment duration has to be interrupted to prevent the seeds to germinate (Lestander 1988).

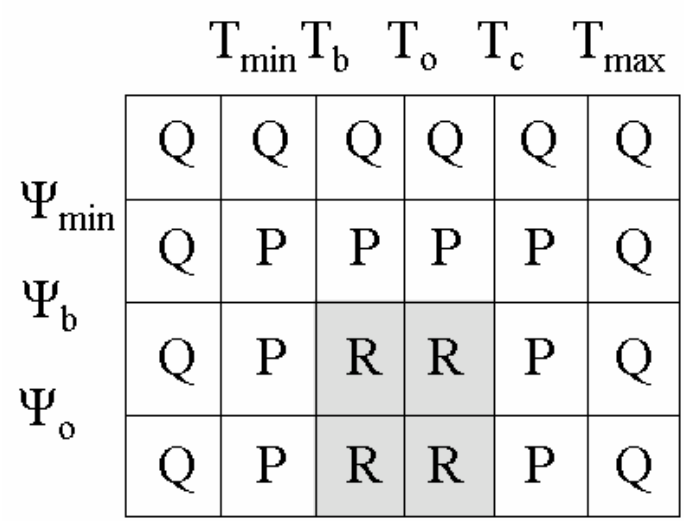

Figure 2. An overview of seed states and their dependence on different combinations of water potential ($\mathbf{\Psi}$) and temperature ($\mathbf{T}$): Q - the seeds become quiescent and there is no progress towards germination; P - seeds progress towards germination but radicle emergence can not occur; R - seeds progress to radicle emergence and germinate. The state Q is defined by $\mathbf{T} < \mathbf{T}_{min}$ or $\mathbf{T} > \mathbf{T}_{max}$ or $\mathbf{\Psi} < \mathbf{\Psi}_{min}(G)$, P by $\mathbf{T}_{min} < \mathbf{T} < \mathbf{T}_b$ or $\mathbf{T}_c < \mathbf{T} < \mathbf{T}_{max}$ and $\mathbf{\Psi}_{min}(G) < \mathbf{\Psi}(i) < \mathbf{\Psi}_o(G)$ and R by the combinations $\mathbf{T}_c > \mathbf{T} > \mathbf{T}_b$ and $\mathbf{\Psi} > \mathbf{\Psi}_b(G)$. The indices are: b for base, o for optimum, c for ceiling, min for minimum and max for maximum.

Stress testing, *e.g.* artificial ageing and controlled deterioration, to measure seed vigour have been developed (reviewed by McDonald 1999). These tests are often conducted for a short time period at elevated moisture and temperature regimes, *i.e.* at temperatures above the ceiling temperature ($\mathbf{T}_c$) and base water potential for germination ($\mathbf{\Psi}_b$). After the stress test a standardised germination test (ISTA 1999) is often carried out and compared to germination performance before the test.

The hydrothermal time of a seed or a seed lot is dependent on its earlier development. Hydrothermal time is related to seed development, dormancy and dormancy loss (Allen & Meyer 2002, Bradford 2002, Batlla & Benech-Arnold 2003) which in its turn are related to embryo osmotic potential (Welbaum *et al.* 1998), dry after-ripening (Allen & Meyer 1998), etc. Experiments have shown that hydrothermal time explains a major part of the variation in germination time. It has also been shown that pre-treatment influences base temperature and base water potential. Rose & Finch-Savage (2003) showed that the base water potential for 50% germination was constant at ca 15 °C near the base temperature for germination (T_b). At temperatures above base temperature ($T>T_b$), the base water potential for 50% germination increased linearly with temperature. When temperature exceeded the optimal temperature ($T>T_o$) the base water potential ($\mathbf{\Psi}_b$) shifted to higher values (Alvarado & Bradford 2002).

The hydrothermal time can be modified as a threshold germination model (Finch-Savage *et al.* 2000) and be used to guide pretreatments based on water and osmotic priming of seeds prior to sowing. One drawback in using this technique is that seeds are either placed on the surface of water solutions of polyethylene glycol (held up by the surface tension) or on germination papers that are in contact with these solutions *e.g.* use of chemical solutions, more handling of seeds etc. Michel (1983) and Hardegree & Emmerich (1990) have developed functions to determine the water potential from -400 MPa to pure water at 0 MPa for such solutions. This technique is difficult and costly to apply at large scale. Depending on the amount of osmotic active solutes in seeds there seems to be no straightforward relation between seed moisture content and water potential. Such relations would make it easier to fully apply the hydrothermal time concept in practice and use so called naked seed hydration (compared to solid matrix hydration) to given moisture contents near the limits for radicle emergence (Thomas 1983, Bergsten 1987). Regulating the water status of biological active seeds in Phase 2 is already applied in large scale for conifer seeds in Sweden (Lestander 1988).

The above reasoning clearly shows that water is a major factor in seed viability management. The seed-water interaction contains information of ongoing biological activity within the seeds and water is under certain circumstances a marker molecule for seed activity and viability. Still the question remains, how to measure the seed-water interaction in a flow of single seeds without causing seed damage?

There are many techniques available to measure the seed-water interaction but methods that use electromagnetic waves are of interest since supporting technical

platforms offer non-invasive and fast measurements. The preferred measurement will therefore be one of scattered radiation from or through seeds.

Electromagnetic radiation

The electromagnetic spectrum extends from the extremely short wavelengths of gamma radiation to the long-range wavelengths of radio waves, Figure 3.

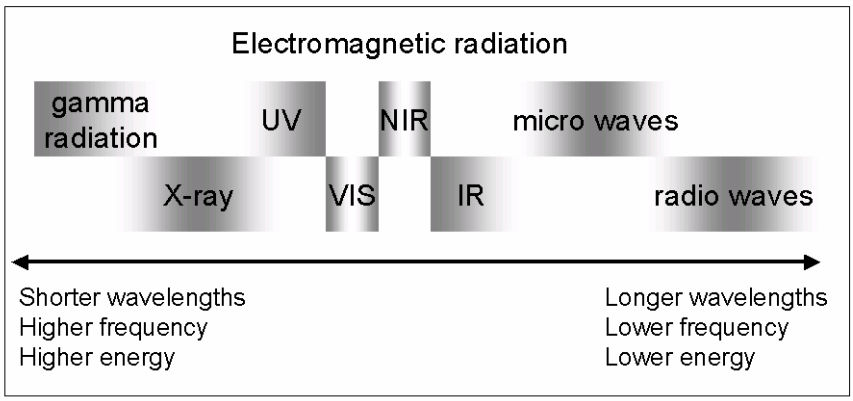

Figure 3. Overview of the electromagnetic wavelengths that range from gamma to radio waves.

According to quantum mechanical theory, electromagnetic waves are radiated when an atomic system shifts from a higher quantum state to a lower one. Energy quanta can also be absorbed when going from a lower state to a higher one.

Gamma radiation is emitted from nuclear reactions and can also cause nuclear reactions when absorbed. X-rays are related to energy levels of inner electrons of the atom. Both gamma and X-ray radiation can therefore be used to detect and quantify atoms. They are also known to cause radiation damage in biological samples. Non-destructive measurement of biomatter is therefore not advised with this high-energy radiation. Water can be used as a contrast agency in the X-ray region and Simak *et al.* (1989) proposed the use of X-ray images for the determination of seed viability in germination tests.

Ultraviolet (UV) and visible radiation interacts with outer atom electrons or crystal structure electrons. UV radiation can also cause ionization. One UV source is radiation emitted from a plasma (Na lamps). An important use of UV radiation for biological systems is detection of fluorescence. One example is fluorescent coating of cabbage where non-viable seeds that leach sinapine give fluorescence at 430-450 nm (Taylor *et al.* 1993). Furthermore, the spectral region is of interest as UV induces fluorescence of key biomolecules (*e.g.* DNA) when tagged by fluorescent probes (*e.g.* Tang *et al.* 2003). The visual (VIS) region (400-780 nm) has been used for a long time for manual sorting of seeds. For nearly half a century automatic colour sorters have been used (Harmond *et al.* 1968). By this technique discoloured seeds showing visible defects are removed, *e.g.* immature seeds,

contaminated seeds etc. The quantum shifts of VIS light are entirely from electron excitation in chromophores. In the visual region chlorophyll fluorescence is used to characterize seeds and plants (Öquist & Wass 1988, Sundblad *et al.* 1990, Jalink *et al.* 1998, Konstantinova *et al.* 2002).

An interesting wavelength region for fast non-invasive and non-destructive measurements of seed-water interaction is near infrared (NIR). NIR radiation interacts with overtones of vibrating bonds in polar molecules (Osborne *et al.* 1993) and penetrates deeper into a sample than UV, VIS or IR. NIR spectroscopy is widely applied in science and industry, mainly in the chemical, pharmaceutical and food industry (Osborne *et al.* 1993, Espinosa *et al.* 1994, Boelens *et al.* 2000, Axon *et al.* 1998, Burns & Ciurczak 2001). It is also one of the fastest growing segments of commercial analytical instrumentation.

In the infrared (IR) part of the electromagnetic spectrum, fundamental vibrations within molecular bonds are found. In far IR rotational absorption occurs. The IR region is excellent for the detection of molecules but a main problem as in UV and VIS is that the radiation mainly interacts with the surface of the samples. Therefore IR is often used in the gas phase. For liquid and solid phases only thin samples can be used and this often requires destructive sample preparation. Information from deeper layers of the sample is more or less concealed depending on the optical density of the sample. This is not a problem when using microwaves or radio waves that have high transmission through most materials. In the microwave region molecular movements are detected. The use of nuclear magnetic resonance (NMR) is based on signals in the region of radio waves to measure nuclear spin energy levels in atomic nuclei of mainly H (hydrogen) and C (carbon) isotopes. Studies in the micro and radio wavelength range have shown promising results in determination of seed moisture content (King *et al.* 1992, Bartley *et al.* 1998, Lawrence *et al.* 1998a and 1998b).

Besides the electromagnetic wavelength region, electrical impedance of seeds has been used to measure seed moisture content (*e.g.* Nelson & Lawrence 1994) and to study seed viability in relation to the seed-water interaction (Repo *et al.* 2002). However the impedance measurement requires that the seeds are in contact with a conductive material, *i.e.* an electrode.

Near infrared spectroscopy

Near infrared (NIR) radiation is in the wavelength range of 780-2500 nm, whereas 400-780 nm is visible (VIS) and above 2500 nm is infrared (IR). The discovery of NIR radiation was done in 1800 by Sir William Herschel (Davis 1990). The only tools he used were a prism that refracted a sunbeam and a thermometer. Beyond the red part of the spectrum he found that temperature rose. Today we know that this measurement was done in the spectral region of molecular vibration.

During the 1930's when infrared spectroscopy was introduced the region below 2500 nm was considered uninteresting and left aside (Norris & Butler 1961). The interest in using NIR spectroscopy on seeds started when Norris and coworkers in

the 1960's found that NIR could be used to determine seed moisture content (Norris & Hart 1996 (reprint from 1965), Ben-Gera & Norris 1968). The commercial breakthrough for NIR spectroscopy came when it was realized that this technique could additionally be used to determine protein content in samples of whole grains (*e.g.* Williams *et al.* 1985). The reason for the popularity of NIR spectroscopy was that since little or no preparation was needed, time, chemicals and thus costs could be saved. In the 1980's, NIR analyses of protein content of grains became an officially approved method in the USA. The history of the NIR technique and its progress can be found in Norris (1988), Davis (1990), Osborne *et al.* (1993), McClure (1994), Hindle (2001, 2002) and Barton (2002).

Today, the NIR technique is widely used not only in chemical, pharmaceutical and food industries (*e.g.* Jedvert *et al.* 1998, Boelens *et al.* 2000, Reich 2002), but also in agricultural and forest industries (Downey 1985 and 1996, Downey *et al.* 1990, Wallbäcks *et al.* 1991, Osborne *et al.* 1993, Thygesen 1994, Antti *et al.* 1996). NIR spectroscopy is also used in environmental studies (Nilsson *et al.* 1996, Geladi *et al.* 1999, Dåbakk *et al.* 2000). Another field of interest is non-invasive clinical diagnostics where NIR spectroscopy has been used to analyze blood, tumours, skin etc (*e.g.* Heise *et al.* 1998, Hazen *et al.* 1998, Hull *et al.* 1999, Geladi *et al.* 2000, Kim *et al.* 2003).

Dowell and co-workers carried out interesting "Russian doll" studies that show the potential of using NIR spectroscopy in seed science (Dowell *et al*. 1998 and 1999, Baker *et al*. 1999). The first step was to detect insects in single grains of wheat. The classification gave 95-96% sorting accuracy. It was also possible to distinguish between insect species. The next step was to detect if the concealed insect had a parasite inside. Even here the sorting efficiency among insect infested seeds was high (90-100%).

NIR spectroscopy studies have shown high potential for the classification of bulk seed samples and single seeds. Examples are fungal contamination in seeds (*e.g.* Hirano *et al.* 1998, Pearson *et al.* 2001), internal insects in wheat (Chambers *et al.* 1992, Ghaedian & Wehling 1997, Dowell *et al*. 1998, Baker *et al*. 1999) or in *Cordia africana* Lam. and Norway spruce (Tigabu 2003, Tigabu & Odén 2002). The NIR spectra also contain information on physical seed properties for example seed weight, seed size, bulk density, etc (Hurburgh *et al.* 1995, Kawamura *et al.* 1998, Velasco *et al.* 1999, Font *et al*. 1999). It has also been proven effective to classify seed viability in a broad sense, such as deteriorated seeds (Soltani 2003) and empty seeds (Tigabu 2003). Varieties of different grains have successfully been classified by NIR spectroscopy (Delwiche & Massie 1996, Kwon & Cho 1998, Turza *et al.* 1998, Delwiche *et al*. 1999) as well as different seed provenances of forest species (Rumler *et al.* 1993).

The main use of NIR spectroscopy within the field of seed science is for quantification of seed moisture content and chemical constituents like protein, oils, etc (Norris & Hart 1965, Ben-Gera & Norris 1968, Halsey 1987, Lamb & Hurburgh 1991, Sato 1994, Campbell *et al.* 1997, Pazdernik *et al.* 1997, Delwiche 1998, Kohel 1998, Sato *et al.* 1998, Velasco *et al.* 1998).

Besides studies showing the potential of using NIR spectroscopy on biomaterials such as seeds, there is a growing interest in building seed sorters using the NIR wavelength region to sort seeds according to different properties (*e.g.* Dowell 1998, Dowell *et al.* 1998, Baker *et al.* 1999, Pearson 1999, Ridgway *et al.* 1999, Pasikatan & Dowell 2001, Dowell *et al.* 2002, McCaig 2002).

Theory and principle

NIR radiation interacts mainly with overtone vibrations of polar molecules (covalent bonds between heavy and light atoms: C-H, O-H and N-H; this kind of notation focuses on vibrating covalent bonds and assumes additional covalent bonds to the C, O and N atom). The energy levels corresponding to fundamental vibrations are found in the infrared region whereas the overtones and combinations of these are found in the NIR. Overtone and combination tone vibration quanta are absorbed or emitted when the vibration energy of a bond is changed.

A molecule with n atoms can be described by a number of $3n$ momenta, also called degrees of freedom (*e.g.* Osborne *et al.* 1993). Three degrees of freedom are used for translation. Such translations exhibit energies in the radio wave and microwave region. Three more degrees of freedom are needed to define the rotations of the molecule giving energy levels in the far IR. Thus $3n$-6 degrees of freedom are left for fundamental vibrations and $3n$-5 for linear molecules such as HCl ($3n$-6 is zero for HCl). Water with its 3 atoms has 3 modes of fundamental vibration in the IR: symmetric stretching, asymmetric stretching and bending. According to quantum mechanics these changes are discontinuous*, i.e.* jumping from one quantum state to another. For H-O-H (water) the symmetric and asymmetric stretching have almost equal energy levels (Efimov 2001, Efimov & Naberukhin 2002) and they can only be separated by high resolution gas phase spectroscopy.

The overtones and combination bands of O-H in water, C-H in carbohydrates and N-H in proteins give high absorptions. Similarly the double bonds of C=O and C=C give absorption in NIR. These bonds are common in biomolecules. Tables of peak absorption with chemical assignments are found in Williams & Norris (1987), Osborne *et al.* (1993) and Shenk *et al.* (2001). Peak location can shift, and is dependent on temperature, interacting molecules and hydrogen bonding. One example is the water peak caused by the first overtone of O-H stretching which shifts from 1491 to 1412 nm when temperature is raised from 6 to 80 °C (Segtnan *et al.* 2001). Water gives high absorbances in the NIR range. Maxima for pure water at 20 °C are at 970, 1190, 1450 and 1940 nm (Curcio & Petty 1951) These water bands are three overtone bands, the first at 1450 nm, the second at 970 nm and the third at 760 nm and bands combining O-H stretching and bending at 1940 and 1190 nm. In this thesis the bands at 1940, 1450 and 1190 nm are called water I, water II and water III, respectively.

The fundamental vibrations are at least a factor 10 stronger in quantum efficiency than the overtones and combination bands (Bokobza 1998). This lower quantum

efficiency is an advantage as it allows deeper penetration of NIR radiation in a sample, thus a higher degree of interaction with deeper layers in the sample. A study using cod tissue has shown that the penetration depth was at least 20 mm (Nord *et al*. 2002). In human skin, NIR radiation penetration is between 0.5 and 2 mm (Marbach 1993). Effective sampling depth in tablets is given as 1.9-2.7 mm for reflection and up to 3 mm for transmission (Iyer *et al*. 2002).

Instrumentation

A spectrometer consists of a radiation source, monochromator, sample cell, detector and readout electronics, Figure 4. In some instruments the monochromator and sample cell exchange places. Some of the rays in Figure 4 can also be replaced by fibre optics. In recent instruments spectral data are automatically saved on a computer hard disc.

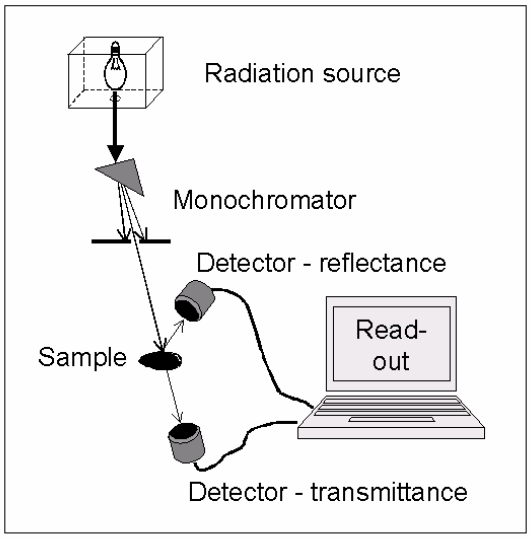

Figure 4. Simplified principle of a NIR instrument.

Sources of NIR radiation

A suitable NIR radiation source is a filament, heated to at least 2500 K. At this temperature the peak radiation of a black body is at 1160 nm. Tungsten halogen lamps or heated xenon gas plasma can be used as sources of NIR radiation as well as tuneable lasers and light emitting diodes (LEDs). Furthermore, synchrotron radiation can be used, but it is an expensive option.

Detectors

One important key to the success of NIR spectroscopy is detector development. Lead sulphide (PbS) detectors were developed in the 1940's and are still the most widely used NIR detector within the range of 1100-2500 nm. The most common detector in the range 360-1050 nm is a silicon detector. An interesting

development is the indium-gallium-arsenide (InGaAs) detector as it is much faster in response than the others. This detector covers the spectral range of 900-1800 nm and has also better sensitivity than the PbS-detector. Most sensors are sensitive to thermal noise and improvements in sensitivity can be made by cooling them.

A problem with the detector is the reading time for development of extremely fast instruments. Today, the fastest NIR detectors have a reading cycle of 0.1-0.01 seconds.

Monochromator setups

NIR instruments types can be classified in many ways. The most frequently used type is based on the monochromator or optical principle. Figure 5 presents different optical setups. Early instruments used selected wavelength bands based on fixed filters. Later full spectrum instrumentation became available. Scanning instruments allow scanning the full spectrum over some time interval and use one or two detectors. Other instruments use many parallel detectors which saves time, allowing fast simultaneous measurements.

	Optical principle	
	Scanning	Parallel
Fixed wavelength	- filter wheel - LEDs - lasers	- sensor modules
Full spectrum	- scanning grating - AOTF - Fourier transform - LED spectrometer - tuneable lasers	- detector array and grating - stationary interferometer

Figure 5. NIR instrumentation based on optical principle.

The instruments used in this thesis (papers I-IV) are all based on scanning with a reflective monochromator grating. The detectors used are silicon up to 1100 nm and PbS between 1100 and 2500 nm. The scanning requires longer measurement time than parallel measurements, especially because a number of scans has to be averaged to reduce noise. Detector arrays remove the need to scan with a grating and can have flexible short integration times. Many manufacturers are moving towards instruments using Fourier transform NIR (FT-NIR) because they can be made very robust and with high spectral resolution. Simple, cheap and fast instruments can be built by having a few parallel detectors with fixed filters (sensor modules, simulated in papers II and III). Some instruments combine the radiation source and monochromator by using LEDs or tuneable lasers. More details of instruments are given in Osborne *et al*. (1993) and Workman & Burns (2001).

The scattering matrix

When matter – for example a seed - is illuminated by an electromagnetic wave the discrete electric charges of the matter, electrons and protons begin to oscillate by the electric field of the incident wave. Secondary radiation (I_s) may occur when accelerated charges radiate electromagnetic energy in all directions. This process is called scattering and it is related to anisotropy in the system of electrical charges. All media except vacuum are anisotropic and thus scatter radiation. This results in phenomena like diffuse reflection by rough surfaces and diffraction by slits, gratings, edges, etc and at optically smooth interfaces specular reflection and refraction (Bohren & Huffman 1998). A part of the incident electromagnetic energy (I_a) may be transformed into other forms (for example thermal energy, fluorescence, etc) if the elementary charges are excited by the incident radiation (I_i). This process is called absorption. The processes of scattering and absorption are mutually dependent (Bohren & Huffman 1998) as $I_i = I_s + I_a$.

$$\begin{bmatrix} I_s \\ Q_s \\ U_s \\ V_s \end{bmatrix} = \begin{bmatrix} S_{11} S_{12} S_{13} S_{14} \\ S_{21} S_{22} S_{23} S_{24} \\ S_{31} S_{32} S_{33} S_{34} \\ S_{41} S_{42} S_{43} S_{44} \end{bmatrix} \begin{bmatrix} I_i \\ Q_i \\ U_i \\ V_i \end{bmatrix}$$

Figure 6. This shows the relationship of the Stokes vector for incoming (I_i Q_i U_i V_i) and scattered (I_s Q_s U_s V_s) radiation and the Mueller matrix consisting of 16 elements.

The electromagnetic radiation is characterized by its Stokes vector $\mathbf{s}_i=[I_i\ Q_i\ U_i\ V_i]^T$ where I is the amplitude, Q and U determine the polarization direction, V is the polarization absolute phase (Stokes 1852) and the sign ^{T} means a transposed vector. When electromagnetic radiation interacts with matter, all the elements of the Stokes vector can be changed into $\mathbf{s}_s=[I_s\ Q_s\ U_s\ V_s]^T$. The relationship between the Stokes vector for incoming and outgoing radiation is given by the Mueller matrix (Mueller 1948) as illustrated in Figure 6. In this thesis only the scattering of S_{11}, the (1,1) element of the Mueller matrix, was studied because of the technical difficulties in measuring the three last elements of the Stokes vector.

NIR measurement modes

Near infrared spectra can be measured in transmission and reflectance mode. Transflectance and interactance modes can be used for measurement of liquids (Kawano 2002). Transmission is rather easy to understand for gases and liquids, but NIR measurements are typically made on solids, emulsions or suspensions of solids in solutions. These materials often do not allow transmission, in which case

then the measurement is made in the reflection mode. The general term diffuse reflection is often used. Instead of reviewing all possible situations in detail it may be useful to look at what happens to a seed.

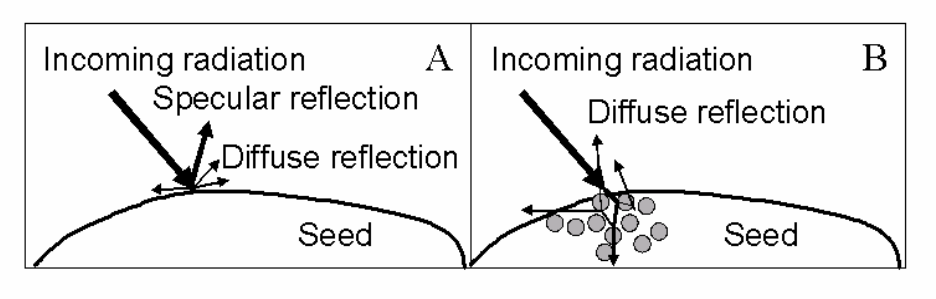

Figure 7. Simplified model of specular (A) and diffuse (A, B) reflection from a seed illuminated by NIR radiation.

Figure 7 illustrates incoming radiation and a seed. The incoming radiation may be a NIR laser beam. Part of the radiation is reflected by specular reflection (Figure 7A). This would happen on a wet seed. The surface acts as a mirror and reflects part of the incoming radiation according to all the laws of mirror reflection. All solids have at least some small percentage of specular reflection and clean and smooth metal surfaces (mirrors) have high specular reflection. Another kind of reflection at the surface is diffuse reflection. A dry seed with its irregular surface would reflect diffusely on the macro scale (Figure 7B). A piece of white velcro or cotton shows a lot of diffuse reflection in the visible region. Diffuse reflection is not always isotropic but the laws of reflection are still followed (Olinger *et al.* 2001). On a microscopic scale Mie scattering and similar phenomena occurs while on a molecular scale there is additional Rayleigh scattering. Mie and Rayleigh scattering are anisotropic and also change the polarization of the radiation (Born & Wolf 1999). All these phenomena represent only different distributions of the reflected light and do not include absorption or changes in energy of the photons. They are however dependent on wavelength, making the situation rather complex when polychromatic light is considered.

Some part of the incoming radiation enters the seed where it can be absorbed or transmitted. (As most pine seeds are black to dark brown, the absorption in visible range is high.) A simplified seed model would be that of a solid particle in a liquid medium of a different refractive index. Transmission into the seed also follows the laws of optics and at each particle boundary the direction of the radiation is changed by reflection and refraction, Figure 7B. Part of the radiation is returned to the surface after multiple refractions and absorptions and this part is measured as reflected radiation. The total ensemble of these processes is called diffuse reflection. There have also been successful attempts to simulate NIR diffuse reflection spectra by the use of the Monte Carlo method (Marbach 1993).

Beer's law

Beer's law, also referred to as the Lambert-Bouguer-Beer law, is the basis of quantitative spectroscopy. This physical law states that the quantity of radiation absorbed by a substance is directly proportional to the concentration of the compound and the path length of the radiation through the substance and can be written as:

$A = \varepsilon cl$

where A is absorbance, ε is the molar extinction coefficient (in L M^{-1} cm^{-1}), l is the path length (in cm) of the sample and c is the molar concentration (in M) of the compound in solution, expressed as M L^{-1}. Thus, the absorbance A has no unit. Beer's law can be rewritten as:

$dI = -I\varepsilon c\, dl$ giving $I=I_i\exp(-\varepsilon lc)$

where dI is the change in intensity of light passing through a substance for an increase in path length of dl. This can be integrated over a given path length and the absorbance is then defined as

$A = \log (I_i/I)$

By setting the incoming radiation to one ($I_i=1$) and relating the collected values of reflectance (R) or transmittance (T) from samples to I_i the absorbance can be calculated as $A=\log(R^{-1})$ and $A=\log (T^{-1})$.

Deviations from Beer's law can occur due to stray light, scattering phenomena or other systematic errors like instrument drift, changes in temperature (Williams *et al.* 1982) during measurement, etc. Such deviations give increased model errors.

Multivariate modelling and regression

The literature on multivariate regression methods is extensive and covers both linear and non-linear approaches (Martens & Næs 1989, Diamantaras & Kung 1996). The problems to be overcome in NIR spectroscopy include the fact that there are many variables relative to the number of observations, but also that the variables are highly collinear, *i.e.* a subset of variables can explain the variance in other variables.

The classical linear regression method is ordinary least squares (OLS) regression, also called multiple linear regression (MLR). The general OLS model is defined for mean-centred data sets as:

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{OLS} + \mathbf{f} \qquad \text{(Eqn. 1)}$$

where **y** is column vector (I×1) of the mean-centred responses (viability, moisture content) for I calibration objects, **X** is the mean-centred matrix (I×K) for I calibration objects (spectra) and K variables (wavelengths), **b** is a vector of OLS regression coefficients (K×1) and finally, **f** is a vector of residuals (I×1).

If the number of observations is lower than the number of variables or if there are collinear variables OLS does not give unique solutions, *i.e.* the solution is not defined. Thus other methods have to be used or the number of variables has to be reduced. One way of reducing variables is to apply principal component analysis (PCA). The obtained result can then replace the original **X**-matrix data and produce principal component regression (PCR), which is a bilinear regression method. It is a two-stage method using first PCA analysis and the OLS on the obtained scores. Another way is the use of partial least squares (PLS) regression that is a generalisation of OLS (Wold *et al.* 1983) and that simultaneously reduces the number of variables and regresses **y** on **X** (Martens & Næs 1989).

The value of the response variable for unknown samples can be predicted by using the regression coefficients in the **b** vector in Eqn 1. For non-centred test sets that usually have other centres than the calibration set these predictions can be calculated as:

$$\mathbf{y}_{pre} = \mathbf{1}y_{mcal} - [\mathbf{X}_t - \mathbf{1}\mathbf{x}_{mcal}]\,\mathbf{b} \qquad \text{(Eqn. 2)}$$

where $\mathbf{y}_{pre}$ is the vector of predictions based on a test set, **1** is a column vector of J ones, y_{mcal} is a scalar and the mean value of the response in the calibration set calculated as $y_{mcal} = (\mathbf{1}^T\mathbf{y})I^{-1}$ with in this case **1** as a column vector of I ones, $\mathbf{X}_t$ is the matrix (dimension J × K) containing the J spectra of a test set, $\mathbf{x}_{mcal}$ is a row vector of K mean spectra in the calibration set calculated as $\mathbf{x}_{mcal} = (\mathbf{1}^T\mathbf{X})\,I^{-1}$ where **1** here is a column vector of I ones and **b** is the vector of coefficients from Eqn. (1).

If the reference values for the test set $\mathbf{y}_t$ are known, a test set residual can be defined:

$$\mathbf{f}_t = \mathbf{y}_t - \mathbf{y}_{pre} \qquad \text{(Eqn. 3)}$$

The residual $\mathbf{f}_t$ can be used in prediction diagnostics, see section Diagnostics.

Principal component analysis

The result of PCA analysis (Jolliffe 1986) is a decomposition of the matrix **X** into informative structure and noise. This is done by maximization of variance directions that are orthogonal to each other and the solution is in the form of a few hyper planes or hyper volumes. The PCA model is often expressed as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \text{(Eqn. 4)}$$

where **T** is a matrix (I×A) of A score vectors (**t**), **P** is a matrix (K×A) of A loading vectors (**p**), the sign ^{T} means a transposed vector or matrix and **E** is the matrix of residuals (I×K). The number A is the rank of the **X** matrix and corresponds to the number of linearly independent rows or columns. The rank is also defined by the number of nonzero singular values. This rank is never used for experimental data and a pseudorank is defined instead. PCA gives orthogonal score vectors in the matrix **T** and an orthonormal basis in the loading matrix **P**. The algorithm uses alternate least squares for finding each component.

The colours of a flag (red and cyan) can be used as an example to illustrate how the PCA decomposes the original **X** data into a hyper plane and how the directions of variance are obtained (Figure 8). The basis colours needed to measure red and cyan are in fact red, green and blue. Cyan is a mixture of green and blue when measured in a spectrometer with the three colour channels blue, green and red. Thus a table of measurements of spots on the flag can be constructed using the three basic colours as variables, Figure 8. This gives a matrix of I×3 values for I observations. The measurement gives values of red colour and the proportions of green and blue in the cyan colour. There are variations between observations both between the flag colours and within a flag colour. The observations can be presented as points in a 3-dimensional coordinate system using the basic colours as the original primary axes as illustrated in Figure 8.

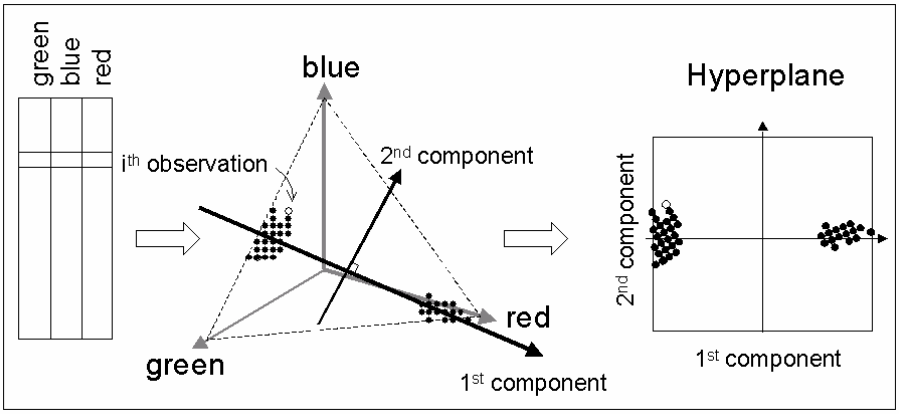

Figure 8. PCA applied on hypothetical spectral data of a flag (red and cyan). The result is a two component PCA model in a hyperplane.

The observed values will fall into a triangular plane with corners in green, blue and red. What PCA does is to find the maximum directions of variance. The largest variation is between the spots of cyan and red. Therefore, the direction of the first component will be between these two colours. This direction is described by the first loading ($\mathbf{p}_1$), called the 1st component in Figure 8. Along this direction the spots get score values. To describe the variance within colours an additional component is need. The direction of this second loading ($\mathbf{p}_2$), called the 2nd component, is where the variation orthogonal to the first component is largest. In this case it is in the direction orthogonal to the first loading, but still in the triangular shaped plane. If the flag had been discoloured more components would be needed to describe the variance among the additional colours.

The result of the PCA is in this case two PCA components that form a hyper plane, Figure 8. The starting matrix (I×3) of I observations of green, blue and red have been reduced to a I×2 matrix by PCA.

Partial least squares regression

The basic equation for partial least squares (PLS) regression is very similar to that for OLS regression using mean-centred data:

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{PLS} + \mathbf{f} \qquad \text{(Eqn. 5)}$$

where **y** is the mean-centred vector (I×1) of the response variable for I calibration objects, **X** is the mean-centred matrix (I×K) for I calibration objects and K variables (wavelengths), **b** is a vector of PLS regression coefficients (K×1) and finally, **f** is a vector of residuals (I×1).

PLS modelling was first described by Herman Wold (Wold 1975, Jöreskog & Wold 1982, Geladi 1988). The orthogonalized PLS regression algorithm developed by Wold (Wold *et al.* 1983) has been reproduced in many studies and described didactically by Antti (1999). The non-orthogonalized PLS algorithm developed by Martens (Martens & Jensen 1983, Martens & Næs 1987, 1989) is presented here. Step 1 and 2 in Table 1 are identical for the two PLS algorithms which give the same coefficients in **b**. Using mean centred or otherwise scaled data sets, the non-orthogonalized PLS algorithm is given as four repeated steps in a sequence given in Table 1. This algorithm is based on three local models that are solved by least squares regression.

Table 1. Steps in the non-orthogonal PLS algorithm

Step	Para-meter	Local model	Least squares solution	Remark
0	**X, y**	$\mathbf{E}_0 = \mathbf{X}$; $\mathbf{f}_0 = \mathbf{y}$		initialization
1	**w**	$\mathbf{E}_0 = \mathbf{f}_0\mathbf{w}_1^T + \mathbf{G}$	$\mathbf{w}_1 = c\ \mathbf{E}_0^T\mathbf{f}_0$	$c=(\mathbf{f}_0^T\mathbf{E}_0\mathbf{E}_0^T\mathbf{f}_0)^{-0.5}$
2	**t**	$\mathbf{E}_0 = \mathbf{t}_1\mathbf{w}_1^T + \mathbf{G}$	$\mathbf{t}_1 = \mathbf{E}_0\mathbf{w}_1$	**G** is a dummy residual
3*	**q**	$\mathbf{f}_0 = \mathbf{T}\mathbf{q} + \mathbf{f}_A$	$\mathbf{q}=(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y}$	$\mathbf{f}_A$ is the residual $\mathbf{T}=[\ \mathbf{t}_1\ \mathbf{t}_2\ ..\ \mathbf{t}_a\ ..\ \mathbf{t}_A\ ]$ $\mathbf{q}=[q_1\ q_2\ ..\ q_a\ q_A]^T$
4	**E** **f**	$\mathbf{E}_1 = \mathbf{E}_0 - \mathbf{t}_1\mathbf{w}_1^T$ $\mathbf{f}_1 = \mathbf{f}_0 - \mathbf{t}_1 q_1$		Go to step 1 with $\mathbf{E}_1$ and $\mathbf{f}_1$. Calculate $\mathbf{w}_2$, $\mathbf{t}_2$ and $\mathbf{q}=[q_1\ q_2]^T$ etc

* **T** and **q** are built up as the algorithm progresses trough more components.

The PLS solution by way of the non-orthogonalized algorithm can also be illustrated as in Figure 9. For each new PLS component that is calculated a new loading vector (**w**) and a new score vector (**t**) are obtained as well as the loading (q) for the reference values after A PLS components have been calculated. The residual for the reference is **f** and **E** for the **X**-matrix. A is the pseudorank of the model.

The coefficients in the **b** vector of Eqn 5 are calculated as:

$$\mathbf{b}_{PLS} = \mathbf{W}\mathbf{q} \qquad \text{(Eqn. 6)}$$

where **W** is the matrix of the loading vectors **w** and **q** is the vector of the loadings q found for the reference variable.

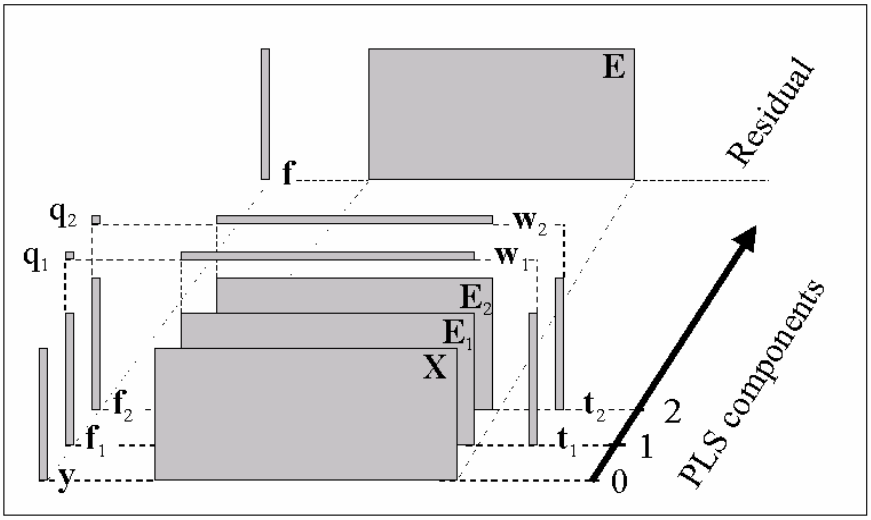

Figure 9. Relations of data and factors in PLS modelling using the non-orthogonalized algorithm.

In PCA the hyperplane is rotated to maximize the explanation of variance in **X**. In PLS each factor is a compromise between maximal correlation to the reference values and maximal explained variance of **X** (Frank 1987, Martens & Næs 1989).

A PLS model is constructed for a number (A) of PLS components. The number A, or the pseudorank, is important. If A is too small, there is underfitting and if A becomes too large, the model explains much of the variation in **y**, but gives bad predictions, a situation of overfitting. Pseudorank is easiest determinated with a test set. The number of components giving minimal prediction residual can be calculated.

When it is considered too difficult to find a test sets, cross-validation can be used. This is done by keeping parts (for example 1/7 of all observations) of the calculation set out as small test sets. When every observation in the calibration set has been out once in such a test set, diagnostics can be calculated. When the test set residual for cross-validation is at a minimum, a good approximation of the pseudorank is found.

In software, for example in SIMCA (Anon. 2000 and 2002d), cross-validation is repeated until every observation has been kept out once and only once to calculate the significance of a new component. By this cross-validation a residual for each observation is calculated into a vector ${}^{cv}\mathbf{f}_a$. The residual of the previous (a-1) components is calculated as $\mathbf{f}_{a-1} = \mathbf{y} - \mathbf{X}\mathbf{b}_{a-1}$. A significant PLS component then has to fulfil the condition that $({}^{cv}\mathbf{f}_a{}^T)({}^{cv}\mathbf{f}_a)[\mathbf{f}_{a-1}{}^T\mathbf{f}_{a-1}]^{-1} < 1$, *i.e.* the residual sum of squares when adding a new component has to be significantly smaller than the previous residual sum of squares ($\mathbf{f}_{a-1}{}^T\mathbf{f}_{a-1}$). For further information see Anon. (2002d).

Bi-orthogonal partial least squares regression

The model interpretation can be done in many ways as PLS offers many parameters and diagnostics. The use of different PLS algorithms also widens the range of parameters to be studied. Therefore bi-orthogonal PLS (BPLS) regression may offer a common platform for interpretation of PLS models as this method unites the solutions of most PLS algorithms. BPLS has the same properties of orthogonal score vectors and orthonormal basis loading vectors found in PCA. The model for BPLS factorisation was described by Ergon (2002), see also paper IV. BPLS is a way of rewriting step 2 of Table 1:

$$\mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{E} = (\mathbf{U}\mathbf{S}\mathbf{V}^T)\mathbf{W}^T + \mathbf{E} = \mathbf{T}_b\mathbf{V}_b^T + \mathbf{E} \qquad \text{(Eqn.7)}$$

where $\mathbf{T}$ is the score matrix (I×A) and $\mathbf{W}$ is the loading matrix (I×K) calculated according to the non-orthogonalized PLS algorithm, $\mathbf{U}$ is an orthogonal matrix (I×A) of eigenvectors ($\mathbf{U}^T\mathbf{U}=\mathbf{I}$, where $\mathbf{I}$ is the identity matrix), $\mathbf{S}$ is a diagonal matrix (A×A) containing the square roots of the eigenvalues of $\mathbf{T}$ (the singular values at the diagonal and off diagonal elements are zero), $\mathbf{V}$ is an orthogonal matrix of eigenvectors (K×A) ($\mathbf{V}\mathbf{V}^T=\mathbf{I}$), $\mathbf{T}_b$ is the orthogonalized score matrix (I×A), $\mathbf{V}_b$ is a matrix (K×A) of orthonormalized loading vectors and $\mathbf{E}$ is the residual matrix (I×K). Here it is assumed that $\mathbf{U}\mathbf{S}\mathbf{V}^T$ has the same rank as the obtained $\mathbf{T}$ and $\mathbf{W}$ matrices.

The BPLS components are found by singular value decomposition of the non-orthogonalized vectors in matrix $\mathbf{T}$. BPLS gives the same solution as PLS when all model components are used. It should however be stressed that the score vectors ($\mathbf{t}_b$) and loading vectors ($\mathbf{v}_b$) are not the same as in the corresponding PLS solution. Further more, the order can be reversed. It is also possible to transform parameters from the orthogonalized PLS algorithm into BPLS (Ergon 2002).

Diagnostics

A number of diagnostics were used in the papers. The coefficient of multiple determination R^2 describes the amount of explained variation in the calibration set (I×K) and is defined as:

$$R^2 = 1 - \mathbf{f}^T\mathbf{f}\,(\mathbf{y}^T\mathbf{y})^{-1} \qquad \text{(Eq. 8)}$$

where $\mathbf{y}^T\mathbf{y}$ is the total sum of squares of the mean-centred $\mathbf{y}$ and $\mathbf{f}^T\mathbf{f}$ is the sum of the squared residuals where $\mathbf{f}$ are from Eqn. 1 or Eqn. 5.

A similar diagnostic using Eqn. 3 to calculate residuals can be defined for the test set (J×K):

$$Q^2 = 1 - \mathbf{f}_t^T\mathbf{f}_t\,(\mathbf{y}_t^T\mathbf{y}_t)^{-1} \qquad \text{(Eqn. 9)}$$

with $\mathbf{y}_t^T\mathbf{y}_t$ is the total sum of squares for the mean-centred $\mathbf{y}_t$ and $\mathbf{f}_t^T\mathbf{f}_t$ is the sum of squares not explained by the model in the test set.

When internal validation (cross-validation) is applied the calculated calibration model can be validated using leave-one-out cross-validation, *i.e.* each observation is left out one time and only once in the cross-validation. The estimated response value (y_{cv}) for the i^{th} observation is then calculated using Eqn. 1 based on the other I-1 observations in the calibration set. The residual in cross-validation is calculated according to Eqns. 2 and 3 giving ${}^{cv}\mathbf{f} = \mathbf{y} - \mathbf{y}_{cv}$. Using this residual the root mean squared error in cross-validation (RMSECV) can be defined and the squared RMSECV is calculated as:

$$RMSECV^2 = I^{-1}({}^{cv}\mathbf{f}^T)({}^{cv}\mathbf{f}) \quad \text{(Eqn. 10)}$$

Other used diagnostics were the root mean square error of estimation (RMSEE) and the root mean square error of prediction (RMSEP). The squares of these are calculated as:

$$RMSEE^2 = \mathbf{f}^T\mathbf{f}\,(I\text{-}A\text{-}1)^{-1} \quad \text{(Eqn. 11)}$$

$$RMSEP^2 = \mathbf{f}_t^T\mathbf{f}_t(J)^{-1} \quad \text{(Eqn. 12)}$$

where I and J are the number of observations within the calibration or test set, respectively, A is the number of PLS or BPLS components in the model. For not mean-centred calibration sets the denominator in Eqn. 11 becomes I-P where P is the number of model parameters including also the offset for example used in OLS regression.

A regression model applied on the test set can be modified to estimate bias, *i.e.* the mean value of the elements in vector $\mathbf{f}_t$ is not zero, using the known coefficients $\mathbf{b}_{OLS}$ or $\mathbf{b}_{PLS}$. Bias is defined as:

$$bias = \mathbf{1}^T\mathbf{f}_t\,(J)^{-1} \quad \text{(Eqn. 13)}$$

where $\mathbf{1}$ is a column vector of J ones (J×1) and J is the number of observations in the test set.

The F-test for comparing different test sets, m and n, was based on:

$$F = RMSEP_m^{\,2}\,RMSEP_n^{-2} \quad \text{(Eqn. 14)}$$

with i and j degrees of freedom equal to the number of tested objects and $RMSEP_m^{\,2} \geq RMSEP_n^{\,2}$, *i.e.* the RMSEP-values as numerator or denominator were chosen so that $F \geq 1$.

Data pretreatment

Regression models are often built as linear regression (Eqns. 1 and 5). This is numerically convenient and also robust, but at the same time a limitation because the true underlying model may be non-linear. An example is Beer's law relating absorbance and concentration. If the NIR data are kept in the transmittance mode,

the model relating spectra and concentrations will definitely be nonlinear, whereas the transformation to absorbance that gives a linear approximation is meaningful. There are also a number of specific techniques for linearizing spectral absorbance data. These are very often based on underlying assumptions for the NIR spectra.

Derivatives of spectra

Derivatives of spectra can be calculated in a number of ways. Usually first or second derivatives are used (Hopkins 2001). In the papers in this thesis, smoothing derivation according to Savitzky and Golay was used (paper I, II and IV). The technique consists of smoothing the data around the wavelength where the derivative is to be found by fitting a polynomial to the point and some of its left and right neighbours. Then the derivatives (first or second) of the polynomial are calculated. The window is moved over the whole spectrum. An example is a window of the point plus 5 left and 5 right neighbours, fitting a 3rd degree polynomial to the 11 points and calculating first or second derivative of the polynomial as a point of the derived spectrum. The basic theory behind taking derivatives is that NIR data often suffer from baselines or sloping baselines. These may be caused by surface roughness or particle size effects, packing effects etc. The first derivative removes the influence of baseline offsets. The second derivative eliminates the influence of a constant sloping baseline. Chemical information is in peaks and these peaks remain in the first and second derivative. They only change shape.

Multiplicative scatter or signal correction

Multiplicative scatter or signal correction (MSC) (Geladi *et al.* 1985) is a technique for removal of baselines and sloping baselines by linear least squares fitting to some reference standard, often a mean spectrum (**m**), with a vector **a** of offsets and a vector **b** of slopes. The vectors **a** and **b** are found by minimizing **E**. They often contain information about particle size, particle shape, packing etc.

$$\mathbf{X} = \mathbf{a}\mathbf{1}^T + \mathbf{b}\mathbf{m}^T + \mathbf{E} \qquad \text{(Eqn. 15)}$$

where **X** is the calibration set of spectral data and with dimension (I×K), **1** is a row vector of ones (K×1), **a** is a vector of offsets (I×1), **b** is a vector of slopes (I×1), and **m** is a vector containing the mean spectrum (1×K).

The MSC corrected spectra are given as:

$$\mathbf{X}_{MSC} = (\mathbf{X} - \mathbf{a}\mathbf{1}^T)[\text{diag}(\mathbf{b})]^{-1} \qquad \text{(Eqn. 16)}$$

where the operator "diag(**b**)" means to put the elements of **b** (Ix1) on the diagonal of **Z** (I×I) (all non-diagonal elements are zero).

This is the same as trying to give all the spectra in **X** a slope of one and a zero baseline. It is assumed that chemical information in peaks has another shape than a baseline or a constant slope and is therefore retained in $\mathbf{X}_{MSC}$.

Standard normal variates

For standard normal variates (SNV) (Barnes *et al.* 1989) autoscaling is done on the transposed matrix $\mathbf{X}^T$ *i.e.* autoscaling of observations by row-wise mean centring and setting the variation among observations to unit variance by row-wise dividing the mean centred values with their standard deviation.

$$\mathbf{c}^T = K^{-1}\mathbf{1}^T\mathbf{X}^T \qquad \text{(Eqn. 17)}$$

where **c** contains the row-wise mean values (I×1) and represents the offset for each spectrum in **X** (I×K), K is the number of variables and **1** is a row vector of ones (K×1).

Row-wise standard deviations (**d**) for the observations are calculated as:

$$\mathbf{d} = [(K-1)^{-1}\ \mathrm{diag}[(\mathbf{X}^T-\mathbf{1c}^T)^T\ (\mathbf{X}^T-\mathbf{1c}^T)\ ]^{1/2} \qquad \text{(Eqn. 18)}$$

where the operator "diag(**Z**)" means to extract the squared diagonal elements of **Z** (I×I) into a vector **d** (I×1) and **1** is in this case a row vector of ones with dimension (K×1). The calculation $\mathbf{X}^T-\mathbf{1c}^T$ (K×I) is used to mean centre rows in the **X** matrix. An element in **d** represents the standard deviation of a row, compensating for the varying slopes in the corresponding observed spectra.

The transposed matrix $\mathbf{X}^T$ is autoscaled by the calculation:

$$\mathbf{X}^T_{SNV} = (\mathbf{X}^T - \mathbf{1c}^T)\ [\mathrm{diag}(\mathbf{d})]^{-1} \qquad \text{(Eqn. 19)}$$

An advantage of SNV over MSC is that no reference spectrum (**m**) is needed.

Orthogonal signal correction

Orthogonal signal correction (OSC) (Wold *et al.* 1998) uses the response variable **y** for pretreatment of spectra before modelling. OSC removes directions that are irrelevant (orthogonal) to **y** out of **X**:

$$\mathbf{X}_{OSC} = \mathbf{X} - \mathbf{X}_{ort} \qquad \text{(Eq. 20)}$$

where $\mathbf{X}_{OSC}$ is used in the PLS equation (Eqn. 5) or in PCA , $\mathbf{X}_{ort}$ is the removed part of the spectral variation. If the split in Eqn. 20 can be made in a robust way, OSC can give an improved and simplified model. It is believed that simplified models give an easier interpretation.

Many ways to calculate $\mathbf{X}_{ort}$ have been proposed (Sjöblom *et al.* 1998, Trygg 2001, Fern 2000, Li *et al.* 2002, Svensson *et al.* 2002). One way to calculate $\mathbf{X}_{ort}$ is by using the algorithm for PCA modified to orthogonalize the score vectors against the variation in the reference variable, *i.e.* to make **T** in Equation 4 orthogonal against **y**. More detailed information of this is given by Anon. (2002d).

Genetic algorithms

Spectral data consisting of up to 1050 wavelengths were collected. As neighbouring wavelength variables in the NIR range are highly collinear, it may be possible to reduce the number of wavelengths for example when regressing spectral data and seed moisture content.

Assuming 1050 original wavelengths and with the goal of selecting only 5 wavelengths, a complete simulation using all combinations results in more than 8.25×10^{12} possible combinations to analyse. More wavelengths give a rapid increase of possible combinations as the number of combinations is $n![(k!)(n-k)!]^{-1}$, where k is the number of selected wavelengths and n is the total number of wavelengths. Due to the huge amount of combinations the selection of predictive wavelengths is an old problem in science. Almost all textbooks on statistics as well as bio- and chemometrics give some method of removing less useful variables (Martens & Næs 1989, Brown 1993, Höskuldsson 1996, Beebe *et al.* 1998, Martens & Martens 2000, Næs *et al*. 2002, Brereton 2003).

The early work on variable reduction in OLS regression was based on forward selection or on backward elimination to find the statistically best suited subset of variables (Draper & Smith 1981). When PCR and PLS regression were introduced attempts were made to use loading plots to select or remove variables (Frank 1987, Lindgren *et al*. 1994, Lindgren *et al.* 1995). Also a variable importance index has been used (Eriksson *et al.* 1999). Newer techniques are the use of jack-knifing (Westad & Martens 2000) to get a standard deviation estimate of the PLS regression coefficients and the use of interval PLS (Nørgaard *et al.* 2000) to find variable intervals of interest.

Genetic algorithms (GA) (Forest 1993, Leardi 2000 and 2001) are considered objective because they do not require a model to be built before the variable selection is done. Genetic algorithms have a stochastic basis. The results of different GA applications can therefore be slightly different. To have more consistent results, the GA process is repeated many times, called programs, to give a more reliable model. However, in the huge amount of possible wavelength combinations a high number of them may give almost the same result.

The terminology used in GA is influenced by genetical science. Variables are called genes and a set of genes (variables) form a chromosome. Genes can mutate and selected chromosomes can pair and recombine. This recombination is called crossover. The variability is obtained by random choice of genes, rate of gene mutations and rate of crossover among selected chromosomes. GA base variable selection on random subsets and improvements of these subsets by simulated genetic evolutionary means.

The exploration of possible variable combinations is done in two steps for PLS models. Firstly, as many genes as variables form sets of chromosomes. Each gene is randomly set as 0 or 1, where 0 means “variable absent” and 1 means “variable present”. A chromosome consists therefore only of zeros and ones. The chromosomes that have the lowest cross-validation errors in PLS prediction are

selected for the next step. In the second step, selected chromosomes are allowed to pair and recombine and genes are allowed to mutate *i.e.* change from 1 to 0 or vice versa. This gives new variable combinations.

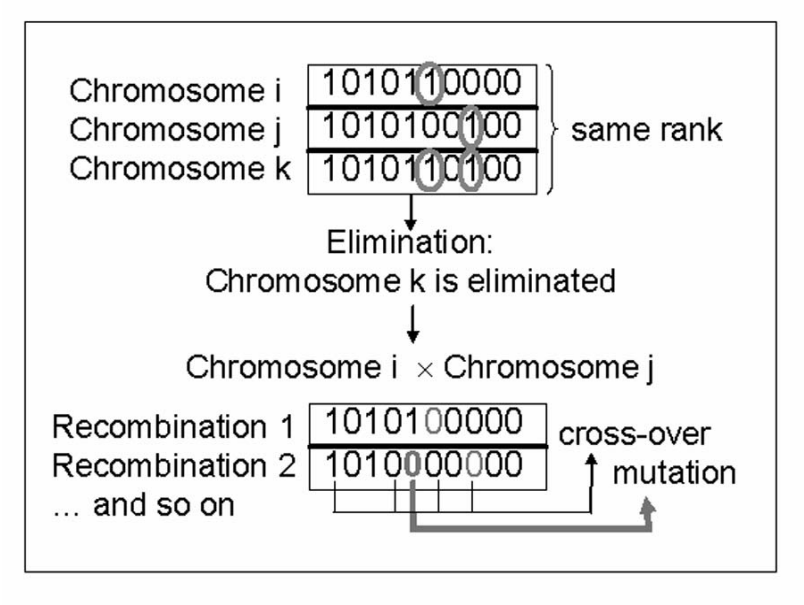

Figure 10. Principle of crossover, mutations and chromosome elimination using genetic algorithms.

Figure 10 illustrates elimination of chromosomes with the same model rank but with more variables present. Also shown is that selected chromosomes recombine randomly according to a preset rate of crossover and how random mutations occur among genes at a preset mutation rate.

The genetic algorithm approach (Leardi & González 1998, Leardi 2000) starts with a population of *n* chromosomes (*n*=30 in paper III). These subsets are tested against a criterion, objectively and quantitatively defining its quality based on the root mean square error in cross-validation (RMSECV) of a PLS model for each subset. Selection of useful variables is based on their frequency of occurrence in the best models obtained for each program.

Data sets with five times more variables than observations should be avoided in GA to lower the risk of random correlation (Leardi 2000). For practical reasons, GA is limited to 200 variables. The number of wavelengths can be compressed by using the mean values of small wavelength intervals. This can be applied when neighbouring wavelengths (variables) are highly collinear. Using such small windows is called iterative GA (iGA). Another way is to use PLS regression on a low number of wide wavelength intervals and to remove those intervals that have low information value. This method is called interval PLS (iPLS) and its

combination with GA (iPLS-GA) is described by Leardi & Nørgaard (2003). A more simple way to tackle the problem of many wavelengths is to divide the spectra into segments of less than 200 variables and do a separate GA on each of them. This is called segmented GA (sGA). After one or more selection rounds by GA, the selected wavelengths are regressed using PLS as in Eqn. 5. Figure 11 shows the frequency of selected wavelengths after the first round of sGA. Prediction accuracy and diagnostics can be calculated using test sets and Eqns. 2 and 3.

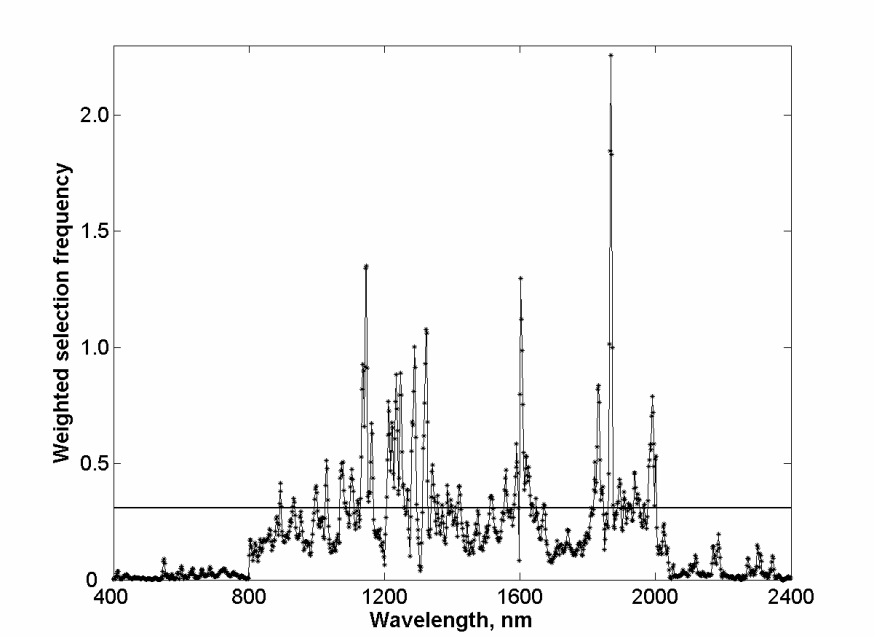

Figure 11. Selected wavelengths (above horizontal line) using segmented GA on single seed data (reproduced from paper III, where the index of weighted selection frequency is given).

The main reason for applying wavelength selection was to test the hypothesis that a low number of NIR sensors can predict moisture content with the same accuracy as using the whole NIR wavelength range. To test this, the selected wavelength bands were simulated as filters in NIR sensors. The filter shape can be different, for example Gaussian filters or uniform density filters. In Gaussian filters the transmission is shaped as the bell curve, *i.e.* maximum transmission in the centre of the interval and gradually lowered to zero transmission at the ends of the interval. A uniform density filter has 100% transmission within the interval and 0% outside the interval, *i.e.* blocks all radiation outside the interval. This simulation compresses the spectrum for each observation into a low number (<10) of absorbance values. By using OLS regression based on these sensor variables the b-coefficients can be calculated according Eqn. 1 and predictions in a test set made according to Eqns. 2 and 3, but also different diagnostics can be calculated to compare different models.

Seed model

In these studies seeds of Scots pine (*Pinus sylvestris* L.) were used as a basis for the seed model. Scots pine is a gymnosperm. The inheritance pattern in seeds of gymnosperms is different from that of angiosperms.

A seed consists of a seed coat, an endosperm - tissue for storing nutrients - and an embryo. The seed coat is entirely a diploid maternal tissue. The embryo is a product of fertilisation between haploid maternal and paternal gametes. In angiosperms the endosperm is a triploid of diploid maternal and haploid paternal contribution. In gymnosperms the endosperm is a female megagametophyte that is haploid (*e.g.* Tillman-Sutela & Kauppio 1995). As a consequence of this about 90% of the seed mass is entirely of maternal origin (Reich *et al.* 1994) making it more easy to classify maternal origin of gymnosperm seeds, but more difficult to characterise paternal origin by spectroscopic means (Lestander *et al.* 2002, Tigabu 2003). Gymnosperms have also a different inheritance pattern of chloroplasts in the embryo. This genetic material comes from the pollinating father (Szmidt *et al.* 1987) whereas in angiosperms the origin is the mother.

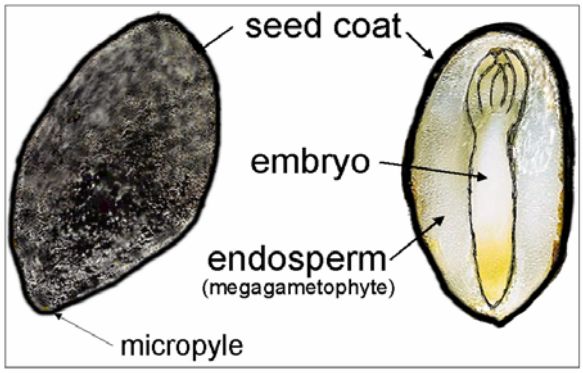

Figure 12. Sketch of a Scots pine seed.

The chemical composition of pine seed is approximately 35 % protein, 48 % oil and 6 % carbohydrate (Bewley & Balck 1994). In a study by Tillman-Sutela *et al.* (1995) of northern Scots pine seeds the oil content varied between 28-33 % of dry weight. The most abundant fatty acid was unsaturated oleic acid.

Pines flower and pollinate during the year after generative bud initiation. The pollen tube growth is arrested after pollination and fertilisation takes place the following summer. In the autumn seeds are fully developed and they are shed the following spring. Thus, the cycle from initiation to mature seeds takes three years. Mean seed weight of Scots pine varies most often between 4 and 7 mg. Single seed sowing is used in nurseries that produce containerized plants. About 70% of the seeds used for forest reproduction in Sweden come from seed orchards (Hannerz *et al.* 2000).

Objectives

The overall objective was to measure seed-water interactions in Scots pine seeds using NIR spectroscopy and multivariate regression models, specially to measure seed-water interaction in single seeds, *e.g.* as a practical application in seed management to enable viability sorting in a flow of single seeds without causing seed damage. The research presented in this thesis covers classification of seed viability, prediction of seed moisture content, selection of wavelengths and interpretation of seed-water interaction modelled and analysed by PCA, OLS, PLS, BPLS, iPLS, wPLS and GA using NIR transmittance and reflectance spectroscopy.

Specific objectives:

- to determine the potential of using NIR spectroscopy to classify viable and non-viable filled seeds (paper I).

- to compare different multivariate regression models for determining moisture content in bulk seed samples and single seeds using NIR reflectance and transmittance spectroscopy, with an emphasis on interpretation of prediction diagnostics (paper II).

- to use genetic algorithms to select wavelengths that could be used in a low number of NIR filter sensors to predict moisture content in bulk seed samples and single seeds (paper III).

- to interpret NIR calibration models based on seed-water interaction of seeds incubated to 30 °Cd (degree-days) and 45 °Cd and modelled by bi-orthogonal partial least squares (BPLS), window PLS (wPLS), second derivatives of raw spectra and difference spectra, but also to study transitions caused by hydration of seeds (paper IV).

The underlying hypothesis was that NIR multivariate calibration models of seed-water interaction can be used for prediction of seed viability and seed moisture content.

Material and methods

An overview of the data sets, instruments, multivariate models and software in the papers I-IV is given in Table 2.

Seeds

The seed samples used in this research had a germination capacity of 0-99% and originated from seed orchards in southern and northern Sweden. All seeds were filled and well-developed. The seeds were incubated to a defined moisture content using controlled hydration described by Bergsten (1987). The seeds were imbibed with deionised Millipore filtered water and incubated as hydro priming. Incubation to 30-105 °Cd (degree-days) was done at 5-15 °C and ca 100% air humidity for 3-7 days before measurement. The incubation cabinet was illuminated by cool white light at ca 20 $\mu Em^{-2}s^{-2}$. Seeds with a moisture content lower than 6% were dehydrated for 0-3 days at 30 °C and ca 30% relative air humidity. In one study (paper I) the seeds incubated to 30% moisture content were dried at ca 25% relative air humidity for 0.8-3.8 h before measurement.

Parallel treatments of seeds provided calibration and test sets. Validations of models were done on test sets.

Reference variables

Seed viability was used in paper I as a model parameter for classification of seed-water interactions by multivariate regression models based on NIR spectroscopy. The seed viability was determined on filled seeds classified as viable or non-viable after a standard germination test at 20 °C (ISTA 1999) and a subsequent cutting test on non-germinated seeds at the end of the germination period (14 days). In some studies, the term non-viable is used in a wide sense. Seeds filled with for example resins as well as empty seeds and badly filled seeds giving low bulk densities are called non-viable. The terms viable and non-viable are used in narrow sense in this thesis, *i.e.* viability of well-cleaned, whole and filled seeds.

Seed moisture content was used in paper II, III and IV as a model parameter for quantification of seed-water interactions using NIR spectroscopy and multivariate modelling. Moisture content (g H_2O g^{-1} fresh weight; in percent) was calculated on fresh weight basis as the weight loss after oven drying at 103 ± 2 °C for 17 ± 1 h (ISTA 1999). Single seed weight was determined by a Mettler MT5 balance with a mass resolution of 10^{-6} g. For bulk seeds a Mettler A30 balance with a resolution of 10^{-4} g was used. In both cases the measurement error was below 0.1% in moisture content. The mean error in single seed moisture content was estimated to 1.3 percent units (absolute value). This was based on weight measurement before and after NIR scanning and was caused by evaporation of water due to low (ca 30% RH) air humidity to which the seeds were exposed during measurements.

Collection of NIR spectra

Reflectance spectra of bulk seeds and single seeds were collected within 400-2498 nm using a NIRSystems 6500 (Silver Spring, MD USA) spectrometer. Each collected spectrum was the mean value of 32 successive scans with 2 nm increments, *i.e.* in total 1050 wavelengths. A spinning cup was used to collect reflectance spectra from samples consisting of 40-50 seeds. For single seed measurement an optical fibre probe was used. These seeds were placed on a black coated metallic surface with negligible reflectance.

Transmittance spectra of single seed were collected four times by using a 1225 Infratec Analyzer (FOSS Tecator, Sweden) with a Single Seed Adapter. This adapter had a sample changer with 23 positions for single seeds. The collected transmittance spectrum for each seed was acquired within 850-1048 nm at 2 nm increments as the mean value of 32 successive scans.

Absorbance (x) was calculated as $x=\log(1/R)$ for reflectance or as $x=\log(1/T)$ for transmittance (T). The underlying assumption was that Beer's law was a good approximation and that inverse calibration models could be used to predict concentrations.

Pretreatments of spectra

Scattering phenomena in solids are wavelength dependent and non-linear. This causes errors in absorbance values. Error correction techniques based on differentiation (first or second derivatives according to Savitzky-Golay differentiation (Hopkins 2001)), multiplicative scatter correction (MSC, Geladi *et al.* 1985), standard normal variate (SNV, Barnes *et al.* 1989) and orthogonal signal correction (OSC, Wold *et al.* 1998) were used. In the reduction of spectral variation according to OSC only one (OSC[1]) or two (OSC[2]) components were used.

In the first study (paper I) 3.4% of the observed seeds were excluded before modelling. The outliers were caused by bad positioning of individual seeds in the single seed adapter of the instrument.

Multivariate modelling

PCA was used in paper I using OSC pretreated data. The model of the viable versus non-viable seed classification in this study was based on a variant of PLS regression based on classes (PLS-DA, Sjöström *et al.* 1986). In all other cases the multivariate model used was PLS or BPLS regression. A Matlab code for converting PLS data into BPLS is given in paper IV. OLS modelling was used on simulated NIR sensors in paper II and III.

The study in paper III aimed at selection of wavelengths that could be used in a low number of NIR filter sensors to predict moisture content in bulk seed samples and single seeds. Genetic algorithms alone or in combination with interval PLS

regression (iPLS, Nørgard *et al.* 2001) were applied to select wavelengths and to test this. A method similar to interval PLS is using a moving wavelength window. This method is called window PLS (wPLS) and was used in study (IV) to find wavelength regions with low loss in prediction or high prediction ability relative to other regions.

Besides these multivariate models, difference spectra and 2nd derivatives of raw spectra were applied also in study IV.

Table 2. Overview of the data sets, instruments, analytical methods and software used in papers I-IV

Studies based on:	Paper			
	I	II	III	IV
Seeds				
- bulk seed		x	x	x
- single seed	x	x	x	x
- test sets	x	x	x	
Instruments				
- NIR transmittance	x	x		
- NIR reflectance		x	x	x
Pretreatment of spectra	x	x		x
Multivariate modelling				
- OLS		x	x	
- PCA	x			
- PLS	x	x	x	x
- BPLS				x
- iPLS			x	
- wPLS				x
- Genetic algorithms			x	
Simulation of sensors		x	x	
Software				
- SIMCA	x	x	x	x
- Unscrambler	x	x		
- Matlab		x	x	x
- PLS_Toolbox				x

Software

Savitsky-Golay smoothing and differentiation was done with Unscrambler (Esbensen 2000, CAMO, Norway) and in PLS_Toolbox (Eigenvector, Manson, USA and Mathworks, USA) for Matlab (Anon. 2003). The spectral pretreatments MSC, SNV and OSC were done using SIMCA 8.0 and 10.0 (Umetrics, Umeå, Sweden). PLS regressions were calculated with SIMCA (Anon. 2002d) and PLS_Toolbox (Wise & Gallagher 1998). Genetic algorithms developed by Leardi (2000) used for wavelength selection were carried out in Matlab. The use of a moving wavelength window in the wPLS approach was based on PLS in PLS_Toolbox and calculated in Matlab.

Results and discussion

Seed viability

The potential of using NIR spectroscopy for sorting viable and non-viable seeds was demonstrated in paper I. This study was based on the finding by Simak (1981) that imbibed viable and non-viable seeds have different drying rates. Filled and well-developed seeds were classified as viable or non-viable after germination. A cutting test of non-germinated seeds at the end of the germination test period was also used to determine viable and non-viable seeds.

The terms viable and non-viable are used in narrow sense in this thesis, *i.e.* viability of well-cleaned, whole and filled seeds. Whereas empty seeds or badly filled seeds giving low bulk densities usually are removed using well-developed techniques. Besides prediction of germination time, prediction of seed viability in a narrow sense may be one of the most difficult tasks in seed sorting. These problems have been tackled by mankind ever since the first farmer sorted seed in the wind.

The problem of classification based on viability is also dependant on the time lag between measurement and endpoint of a viability test *e.g.* tetrazolium staining (Hampton & TeKrony 1995) or a standard germination test (ISTA 1999). If a germination test is used the time lag can be weeks. Thus, seeds can show signs of being viable at the time of measurement, but be classified as non-viable at the end of a test period. In the most extreme case all tissues or cells in a seed may be classified as viable by some means, but deterioration processes inside a seed may develop so far that it will be classified as a non-viable seed at the end of the germination test. There are also cases in which only parts of a seed are dead, *e.g.* frost damaged tip of the radicle, which prevents germination. The problem of sorting seeds into viability classes without overlap is complicated and difficult to solve.

Figure 13 presents the effect of different sorting accuracies. Using a random choice, *i.e.* a sorting accuracy of 50%, no improvement in seed viability is obtained (Figure 13). An assumed target for applying single seed sowing is at least 95% viability. This can than be obtained by sorting seed lots with 13% or lower amount of non-viable seeds by methods with 75% sorting accuracy. By increasing the sorting accuracy from 90 to 99.9% it is theoretically possible to achieve 95% viable seed from seed lots containing 33 to ca 97% non-viable seeds, respectively (Figure 13). Therefore, continued research concerning fast, non-invasive and non-destructive sorting methods to improve seed viability is of great interest as it saves resources.

The sorting accuracy for viable and non-viable seeds in the first study (paper I) was 98-99% using NIR transmittance of single seeds within the 850-1048 nm spectral range, *i.e.* the hypothesis was confirmed. This is extremely high and may reflect the experimental conditions of using killed seeds classified as non-viable seeds by the PLS models. Each seed class used in paper I was uniform. In natural

seed lots there are several background factors to why seeds are non-viable. To overcome this, it may be necessary to construct several calibration models to remove non-viable seeds caused by frost damages, heat damages, mechanical damages etc.

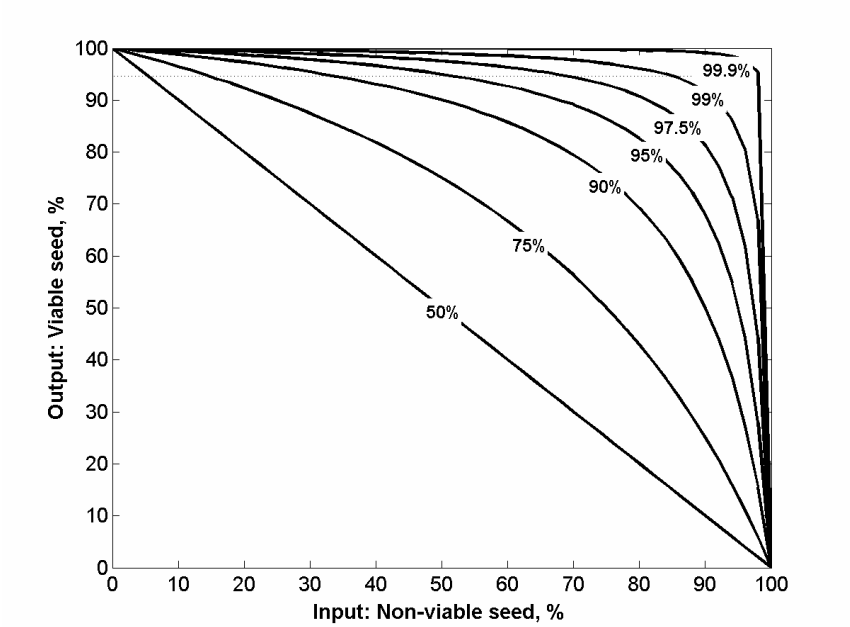

Figure 13. Effect of increased sorting accuracy by removal of non-viable seeds.

The separation of seed viability classes in paper I was based on shifts in seed-water interaction when imbibed seeds were dried. The possibility of using water as a marker molecule for viability under certain conditions was confirmed by the peak at ca 970 nm of the PLS models in paper I using NIR transmittance spectroscopy. This peak is the second overtone of O-H stretching and is assigned as water (Osborne *et al.* 1993).

The method used for controlled hydration of seeds may give higher variation in moisture content between single seeds than using osmotic priming where single seeds are in direct contact with a solution of given water potential. A minor study (data not shown) of single seeds incubated as controlled hydration to 30% moisture content showed that the coefficient of variation was 16.1%. Such starting differences in single seed moisture content most probably increases the calibration model errors. However, the coefficient of variation increased as expected to 30.1 and 40.0% after 2 and 4 h of drying, respectively, due to both increased standard deviation and lowered mean using a seed lot with 65% viable and 35% non-viable seeds. The result in paper I indicated less variation among non-viable seeds with increasing drying period, whereas the variation within viable seeds was not influenced by the drying time.

Seed moisture content

NIR reflectance (780-2280 nm) and transmittance (850-1048 nm) spectroscopy were in paper II used to predict seed moisture content in bulk seed samples and single seeds. Different conditions were simulated based on bulk samples and single seeds using different NIR wavelength ranges and different pretreatments of spectra in the models. Transmittance and reflectance measurements were compared within 850-1048 nm.

In total 34 PLS models were constructed which contained from 1 to 9 PLS components. The errors within the calibration set (RMSEE) ranged from 0.6 to 3.2%. The explained variation of seed moisture content in the test set was within the interval of 76 to 99% and the prediction error (RMSEP) varied between 0.8-3.5% with a bias of up to 2.1%. The overall result from the study in paper II was that NIR spectroscopy in combination with PLS regression provides good models for seed moisture determination.

The more information that is added to a model the better it will perform, provided that the added information contains additional information structure and has low levels of noise and errors that have limited influence on the information structure. The results in paper II showed that the single seed models were improved and more influenced by expanding the wavelength interval from 850-1048 nm to 780-2280 nm than bulk seed models. This change was significant for all average parameter values of single seed models. This can be an artefact of how the NIR reflectance spectra were obtained. Single seeds were measured by using a fibre optic probe that integrated the signal from just a part of the seed, whereas bulk seeds were measured in a rotating cup that integrated the NIR signal from 40-50 seeds. Seed size influenced the measurement distance as single seeds were placed on a black coated metallic surface at a fixed distance to the measurement probe. This contributed to the larger variation in single seed spectra than in bulk seed. Figure 14 shows the standard deviation of absorbance within 780-2280 nm for single seed and bulk seed data. The single seed models used more PLS components than bulk seed models to compensate for this increased variation.

The unexplained variation in predicted seed moisture content increased in the following order:

- bulk sample reflectance at 780-2280 nm or 850-1048 nm,
- single seed reflectance at 780-2280 nm,
- single seed normalised transmittance at 850-1048 nm,
- single seed reflectance at 850-1048 nm,
- single seed transmittance at 850-1048 nm.

Reflectance measurement of bulk samples generated the best fitting models with a minimum prediction error (RMSEP) of 0.8%. The corresponding PLS model explained 99% of the variance in the test set reference. This model is presented in Figure 15.

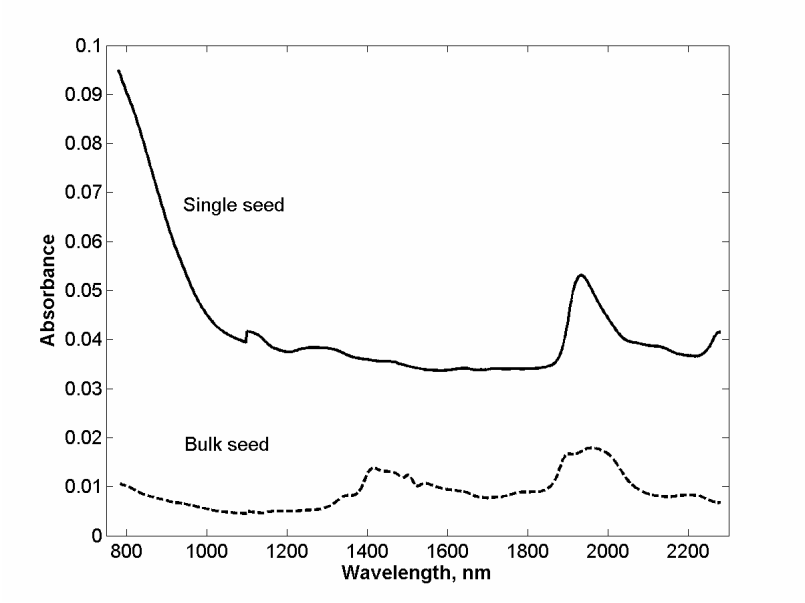

Figure 14. Standard deviation of absorbance values of single seed and bulk seed reflectance spectra within 780-2280 nm.

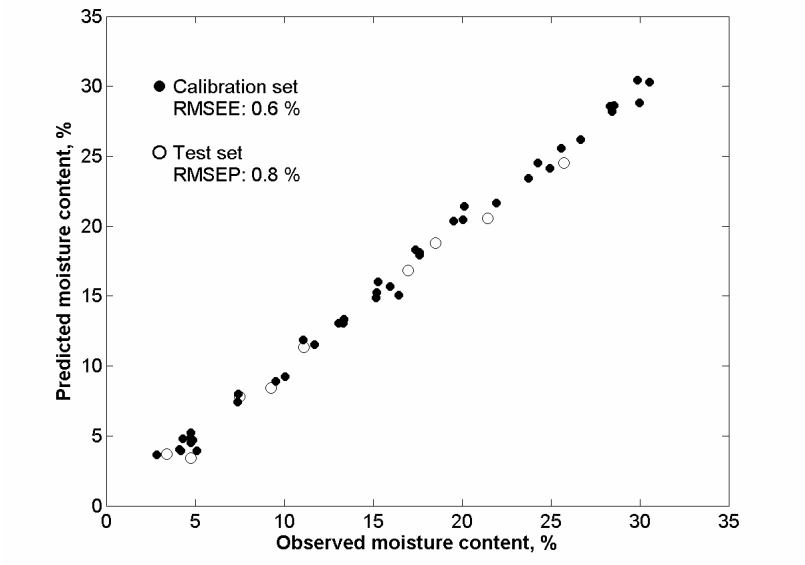

Figure 15. Observed and predicted seed moisture content in calibration (filled circles) and test set (unfilled circles) using a one PLS component model of OSC pretreated spectra (780-2280 nm) from bulk seed samples.

The best single seed model, presented in Figure 16, explained 93% of the variation in the test set and had a prediction error (RMSEP) of 1.9%. This error was larger than for bulk seed. A part of the larger error is caused by an increased error in the determination of the reference variable values. Single seed weights before and after NIR measurement indicated a mean error of 1.3 percent units (paper II). This illustrates the problem of using less well-defined reference values in calibration models. The prediction can never give smaller errors than the resolution in reference determination, although NIR spectroscopy gives high accuracy in measurement of spectral values. For wide seed moisture intervals (between ca 5-32 %) it is difficult to make exact weight determinations of single seeds. This is accentuated for small single seeds (ca 4-12 mg) as they expose a large surface area relative to their mass. For most of the seeds in such wide moisture series water is either evaporated or absorbed depending on the surrounding air humidity at the time of measurement.

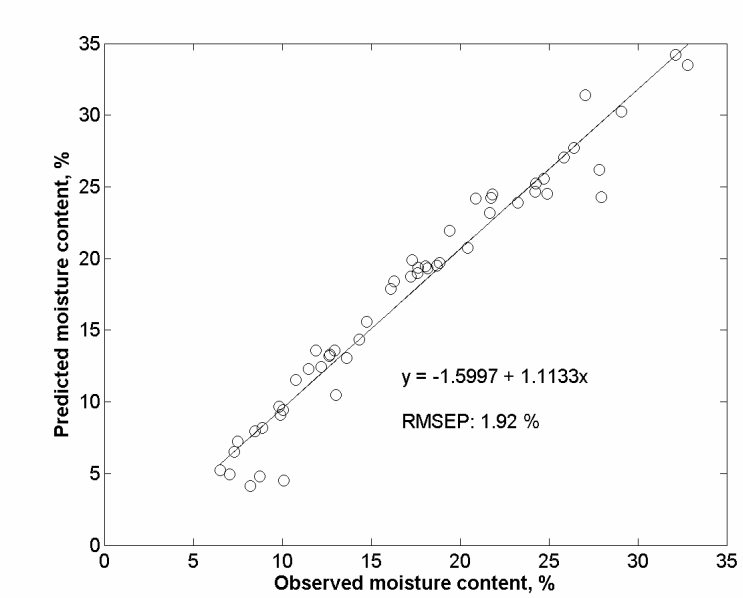

Figure 16. Observed and predicted seed moisture content in test set using a 9 PLS component model based on raw spectra (780-2280 nm) from single seeds.

The best single seed model for NIR transmittance within 850-1048 nm gave an RMSEP of 2.7% and explained 86% of the variation in the test set. The transmittance models used fewer PLS components and had lower bias than those based on single seed reflectance within 850-1048 nm.

The spectral data pretreatments using derivatives, MSC, SNV or OSC (paper II) gave significantly lower errors only in a few cases. This stresses the importance of obtaining spectral data with as little error as possible.

In conclusion, the hypothesis was confirmed and NIR spectroscopy provided good PLS models for seed moisture determination that can be used in forest seed management, both for seed samples and single seeds.

Wavelength selection and simulation of sensors

The OLS simulation of 4-sensors NIR systems in paper II gave prediction errors (RMSEPs) ranging from 3.4 to 4.9 % in the test set. These simulations explained only 52-76% of the variation in the test set indicating that these systems may not be optimal for determination of single seed moisture content. Sensor systems of that resolution can be used only for a rough and overlapping classification of a few seed moisture classes, but not for accurate and precise measurements of moisture content in single seeds. The PLS models predicting single seed moisture content in paper II indicated that 6-9 components were needed to achieve RMSEPs of 1.9-2.2%. This was also an indication that there were 6-9 phenomena in the NIR spectra that contributed to the prediction accuracy. Thus, an increased number of well-chosen sensors would provide for better models.

In paper III, genetic algorithms (GA) were used to select wavelength regions within 780-2280 nm. The result gave good prediction models for seed moisture determination. The RMSEPs of the final models were not significantly different from that of the PLS models using all NIR wavelengths within 780-2280 nm.

Two GA selection rounds or only one in combination with interval PLS (iPLS) resulted in 3-8 wavelength bands with 34-178 wavelengths. The three variants of GA (iPLS-GA, iterative GA and segmented GA) left out the visual region and wavelengths above 2194 nm. The selected bands chosen did not always overlap between the different GA approaches due to the stochastic basis of the GAs. NIR is known for its broad absorption bands and wavelength bands chosen within a 10-50 nm range most probably indicate the same phenomena in the spectra, especially when not peaks are involved.

Figure 17 illustrates the final selected regions. Most of the selected regions were within 1140-1650 nm and located around the slopes and the peak of the water III band (ca 1190 nm, combination band of O-H stretching and bending) and at the slopes of water II (the first overtone of O-H stretching at 1450 nm). Only six regions out of 30 were outside this wavelength interval. An interesting observation was that the wPLS regions in paper IV classified as having a high ability to predict moisture content in many cases overlapped the GA selected regions in paper III.

A few of the GA selected regions were subdivided giving 6-8 final regions that were used to simulate uniform density filters. This kind of filter has equal high transmission at wavelengths within selected regions but block all other wavelengths. Other filter shapes, *e.g.* Gaussian filters as used in paper II, could also be used in these simulations. To simulate sensors equipped with uniform density filters the mean absorbance of each GA selected region was calculated for each observation. In other words the selected wavelengths were further compressed into 6-8 average absorbance values for each observation. The result of the OLS models based on these mean absorbance values explained 89-99 % of the test set variation and the RMSEP values ranged from 0.7% to 2.3 % for prediction of seed moisture content in test sets. These results were of about the same quality as those for the whole NIR range of 780-2280 nm.

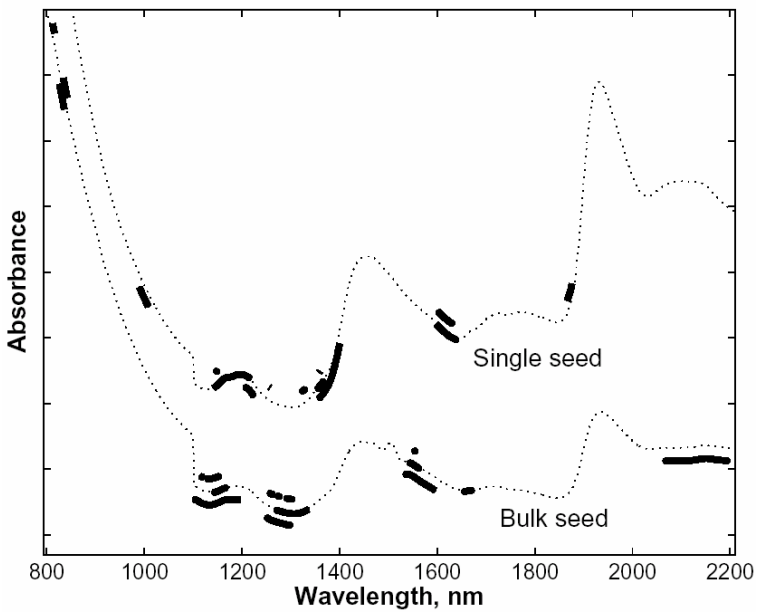

Figure 17. Selection of wavelength regions using GA. The thin line is the mean absorbance between 800-2200 nm for single seed and bulk seed, respectively. Bold lines indicate selected regions according to interval-PLS-GA (underneath the line of mean absorbance), iterative GA (in line) and segmented GA (above line). The mean absorbance lines are shifted horizontally. (Reproduced from paper III).

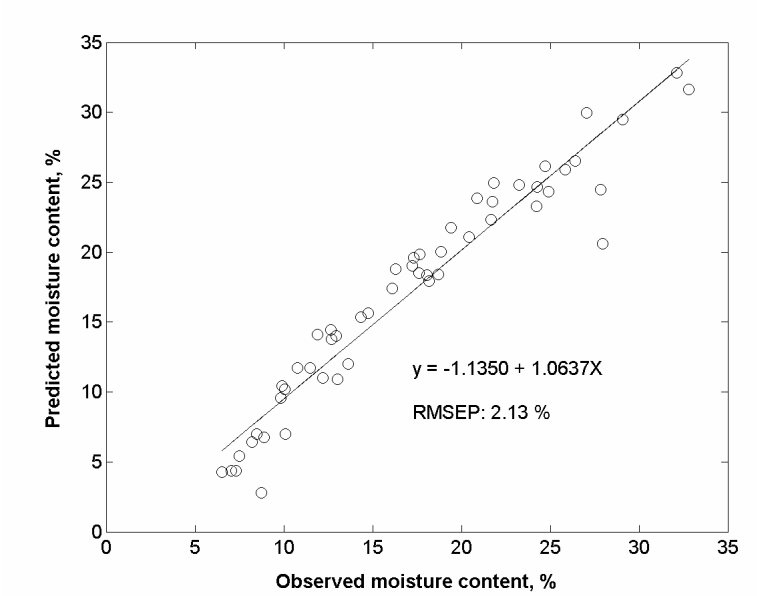

Figure 18. OLS model for simulation of six uniform density filter selected by iPLS-GA for the prediction of moisture content in single seeds.

The best OLS model for single seeds, shown in Figure 18, was based on six simulated uniform density filters selected by iPLS-GA. This gave an RMSEP of

2.1 % and explained 91 % of the variance in the test set. The result was not improved by adding 2-factor interactions, square roots or squares of the filter values. The best result was based on a first order polynomial modelled by OLS according to Eqns. 1 and 2.

It was possible to show that the wavelength bands chosen for the single seeds also worked for the bulk samples and vice versa. This opens a possibility to make calibration models based on bulk reference values instead of single seeds and to avoid the difficulties of obtaining error-free determinations of single seed moisture content.

As high speed sorters using the NIR wavelength region are greatly needed, the results of the GA selection of wavelengths were encouraging. Seed sorters with high throughput, that use only a limited number of NIR bands in combination with present technical platforms of colour sorters, may be developed for fast and non-destructive determination of seed-water interactions.

Due to the need for high throughput of seeds all calculation times should be brought to a minimum and it may be necessary to calculate prediction values directly from the detector signal that is not transformed to absorbance. This would require non-linear regression modelling. It has also been suggested to construct optical filters that compensate for the transformation to absorbance and simulate the b-coefficients in PLS modelling, *i.e.* the detector value is **xb** where **x** is the spectrum from an observed sample or single seed. If such filters can be made available there will be no need for any calculations as the detector value itself is the prediction of the concentration or the classification.

Seed-water interaction

When measuring moisture content in organic matter the peak wavelengths of water are often used in calibration models. The results in this thesis (paper III and IV) show that this intuitive notion does not seem to hold for seeds. One reason for this is that seeds can be considered as mixtures of water and dry matter. Making a model for concentration of water in seeds automatically means making a model for seed dry matter. Another reason is that seeds become more biological active when moisture content is increased. Table 3 gives an overview of important wavelength regions for the models described in paper III and IV. These were found by analysis of BPLS loadings and by studying wavelengths important for prediction of seed moisture content as suggested by GA and wPLS. The results for 1100-2200 nm in Table 3 are from paper III and IV based on reflectance spectra from 30 °Cd single seeds and 45 °Cd bulk seeds.

The most astonishing result concerning seed-water interaction was that, with only a few exceptions, the main water peaks at around 1930-1940 nm (water I), 1450 nm (water II) and 1190 nm (water III) were not selected when using GA and wPLS. It then remains to explain why the other, less prominent, wavelength regions were included. In the following analysis overtone vibrations of covalent

bonds are taken into consideration and they are denoted as for example O-H, C-H, N-H and C=C. This way of notation, as stated earlier, assumes additional covalent bonds to the O, C and N atoms, for example H-O-H, R-C-H and R-N-H.

An important region of the spectrum was 1100-1194 nm (Table 3). This region was found in both single and bulk seeds and was narrowed down to 1126-1170 nm. The b-coefficients of the BPLS models indicate small peaks at 1150 and 1160 nm. The assumed chemical assignment is overtone vibrations of C-H and of C=O double bonds. The C-H vibrations occur at 1143 nm in aromatic rings and at 1152 nm in $-CH_3$ groups. The double bond vibration of C=O is located at 1160 nm (Osborne *et al.* 1993, Shenk *et al.* 2001). Finally, C-H vibrations occur at 1170 nm in structures containing HC=CH such as unsaturated fatty acids (lipids). According to Tillman-Sutela *et al.* (1995), the most abundant unsaturated fatty acids in northern Scots pine seeds are triacylglycerols containing 54 acyl carbons and 5-7 double bonds in the acyl chains. The C=O double bonds are frequently found in fatty acids and their hydrolysis and oxidation.

Due to the wide range both in moisture content and degree-days this narrow region of 1126-1170 nm is most probably also associated to seed respiration. It is well known that moisture content regulates the respiration rate in seeds (Bewley & Black 1994). Double bounds between C and O are frequently found in the tricarboxylic acid cycle (TCA). TCA related enzyme activity in seeds evolves at elevated moisture levels (Falk et al. 1998, Shen 2000, Logan et al. 2001, Benamar et al. 2003).

In the region 1200-1400 nm mainly C-H overtones of the functional groups of –CH, $-CH_2$ and $-CH_3$ are found (Shenk *et al.* 2001). All the GA variants used selected in 1200-1400 nm a total of 11 wavelength bands out of 26 within the 1100-2200 nm spectral range. This indicated many regions with high linear response to increased moisture content. A plausible explanation is that this region also was influenced by catabolism but may also reflect the complementary model of carbohydrates in dry matter of seeds.

The region of 1400-1600 nm contains the water II peak. Three bands at 1532-1594 nm were selected by GA. There were also two major peaks in the BPLS loadings. The first was in 45 °Cd seeds (bulk seeds) at 1416 nm indicating absorption in $–CH_2$ or in aromatic ring C-H structures. The second in the 30 °Cd seeds (single seeds) was at 1460 nm assigned to the first overtone of N-H stretching (Osborne *et al.* 1993). As found in study IV it seemed that this absorption in N-H stretching was changed from 1460 nm to 1502-1506 nm and 1960-1964 nm for the more developed seed states in the 45 °Cd seeds. This may indicate continuing protein de-folding and raised protein metabolism.

At 1600-1800 nm three bands were selected in the interval of 1596-1658 nm in the 30 °Cd single seeds using GA. One GA-band at 1652-1670 nm and two wPLS bands were selected in the 45 °Cd bulk seeds. The model loadings showed peaks at 1632-1636 nm, 1688-1696 nm and 1718-1722 nm. The latter peak can be assigned to vibration in the carbohydrate group $–CH_3$.

Table 3. Overview of wavelength regions with major peaks in BPLS models and regions selected by genetic algorithms (GA) or window PLS (wPLS) within 1100-2200 nm

Peaks	**Single seeds**		**Bulk seeds**	
	GA	wPLS	GA	wPLS
1100-1200 nm				
1144-1152 (C-H)	1140-1214 b 1144-1148 c	1136-1164 (C-H, C=O)	1100-1194 a 1112-1118 c 1126-1152 c 1140-1166 b	1140-1170 (C-H, C=O)
1200-1400 nm				
1204-1206 1258-1266 1340-1368 (C-H) 1396-1406 (C-H)	---1214 b 1204-1222 a 1254-1256 c 1320-1328 c 1352-1366 b 1356-1358 c 1356-1398 a	1380-1414 (O-H, C-H) 1376-1530* (water II, O-H, C-H, N-H)	1248-1298 a 1254-1262 c 1268-1334 b 1270-1276 c 1288-1302 c	1276-1322 1386-1414 (O-H, C-H) 1398-1554* (water II, O-H, N-H, C-H, C=O)
1400-1600 nm				
1416 (C-H) B 1460 (N-H) S 1502-1506 (N-H) B 1544 (O-H) B 1584-1586 (O-H)	1596-1638 b	--- 1530* (water II, O-H, C-H, N-H) 1568-1576 OSC (N-H)	1532-1594 a 1540-1562 b 1550-1554 c	--- 1554* (water II, O-H, N-H, C-H, C=O)
1600-1800 nm				
1632-1636 S 1688-1696 (C-H) 1718-1722 (C-H)	--- 1638 b 1600-1630 c 1644-1658 a		1652-1670 b	1652-1664 1764-1772
1800-2000 nm				
1876-1902 (C=O) 1926-1940 (O-H) 1960-1964 B (N-H, O-H)	1864-1874 c	1852-1882 1846-1890*		1840-1872 1852-1892*
2000-2200 nm				
2050-2080 (N-H) 2124-2140 (C=O etc)			2064-2194 a	2026-2140* (N-H) 2126-2134 (C-O, C=O)

B: Bulk seeds; S: Single seeds; a: iPLS-GA; b: iterative GA; c: segmented GA;
* OSC pretreated data sets

The interval of 1800-2000 nm was dominated by the water I peak at ca 1930-1940 nm. Major peaks in BPLS loadings occurred in the interval of 1876-1902 nm occurring on the leading slope of the water I peak. Although no chemical assignments were found, except the C=O vibrations in functional groups of -COOH at 1900 nm, this slope may contain interesting information in viable seeds.

One reason for this was that wPLS selected this region for both seed sets and sGA selected one region in single seeds.

The 2000-2200 nm region contained major peaks in BPLS loadings suggesting contributions from O-H, N-H, C-H and C=O vibrations. This was further indicated in the 45 °Cd seeds as the region of 2026-2194 nm was selected by iPLS-GA and wPLS in these seeds.

In conclusion, the main difference between the two data sets was within 1100-2200 nm a more complicated pattern of the 45 °Cd bulk seeds compared to the 30 °Cd single seeds. This was most probably an effect of increased hydrolysis, seed respiration and protein metabolism, perhaps also involving DNA synthesis.

Three major transitions in NIR spectra of seeds were found by analysis of raw and 2nd derivatives versions of spectra. These were situated at ca 6, 16 and 27 % moisture content, respectively, and were more pronounced in bulk seed data incubated at 45 ºCd than in the 30 ºCd single seeds. The transition at 6 % was an artefact of the experiment, *i.e.* seeds with moisture content lower than 6% were taken directly from storage conditions or dried before measurement. The transition at ca 16 % was most likely related to the occurrence of free water. The underlying variation also showed a plateau at 16-26 % moisture content. This plateau could be associated with the re-arrangement of membrane configuration as this process needs higher than 25 % hydration of seed dry weight to stabilize (Bewley & Black 1994). Finally, the transition at 27 % may be a combination of increased respiration and protein metabolism. This underlying variation of transitions at small wavelengths intervals implies that local models based on intervals between two transitions, instead of a global calibration model for the whole seed moisture range, may improve prediction accuracy.

A difficulty in interpreting NIR spectra when using varying moisture contents is that hydrogen bonding occurs and causes peak shifts. Vibrations with a high degree of stretching are more affected by hydrogen bonding than bending vibrations. Increased hydrogen bonding tends to slow down the frequency of the vibrations, *i.e.* the bonds become more rigid. The effect is a wavelength shift of peak absorbance to longer wavelengths. This was also demonstrated in paper IV. The water peak at ca 1450 nm showed a shift to longer wavelengths when moisture content was increased. This shift was due to the effect of lower hydrogen bonding as that peak is based on stretching O-H vibrations.

Another problem occurs in the interpretation of the broad and overlapping vibration bands. There are multiple ongoing processes in viable non-dormant seeds when moisture is raised at otherwise favourable conditions for germination. A popular description of this could be: It's like scanning a city from an airplane and based on the scanned information trying to understand what's going on. But also in a city there are major processes. The main processes in seeds are enzymatic activation, increased hydrolysis and respiration, transitions in membranes, protein de-folding and increased protein and DNA synthesis as moisture content is raised from a low level (*e.g.* Bewley & Black 1994, Alberts *et al.* 1994, Kigel & Galili 1995, Copeland & McDonald 2001). As NIR radiation mainly interacts with

hydrogen covalent bonds of polar molecules no definite answers are given, but spectral profiles suggest structural groups of interest for analysis by other means. Thus, NIR spectroscopy cannot give the desired answers alone and scientific cooperation over a wide range of fields is necessary to better interpret NIR spectra using different analytical tools.

Due to continuously ongoing processes in seeds the measurement of seed-water interaction is performed on a non-steady state system. The NIR measurements are often conducted in an environment most suitable for the instrument, *e.g.* at low air humidity, which increases the difficulty in obtaining measurements at equilibrium.

An additional problem is the use of different PLS algorithms that widens the range of parameters to be studied. This was tackled by using bi-orthogonal PLS (BPLS) that produces orthogonal scores and orthonormal loadings and can be used as a common platform for PLS based interpretation of NIR spectra.

Further studies using NIR would include improved spectral resolution (FT NIR) and removal of parts of seeds by mechanical means. Dissolution of seed components and analysis by LC-MS, GC-MS, MS-MS, electrophoresis, etc could give information about the presence of structural elements. Other spectroscopic techniques could be used to give complementary information: FT IR, Raman, solid state NMR, etc.

Conclusions

The NIR technique offers large possibilities to measure seed-water interactions in seeds. As shown within this thesis seed-water interaction involves changes in mainly C-H, C=O and N-H bonds emanating from ongoing biological processes like seed respiration and protein metabolism. Thus, measurement of seed moisture content must also target these overtones bands in addition to the water overtone bands.

The potential of using multivariate NIR calibration models for classification was demonstrated using filled viable and non-viable seeds that could be separated with low class overlap.

It was also shown that multivariate NIR calibration models gave low errors in prediction of seed water content for bulk seed and single seeds using either NIR reflectance or NIR transmittance spectroscopy.

Genetic algorithms selected three to eight wavelength bands in the NIR region and these narrow bands gave about the same PLS prediction of seed moisture content as using the whole NIR interval in the calibration models. The OLS result of simulations of 6-8 NIR sensors based on selected wavelength bands had the same good prediction of seed moisture content. This finding offers good possibilities to apply NIR sensors within present technical platforms for colour sorting.

As the NIR region contains overlapping wide bands from vibrating bonds of polar molecules and as data obtained from NIR instruments are collinear multivariate models like PLS regression must be used to obtain good prediction models. Bi-orthogonal partial least squares (BPLS), which has orthonormal loadings and orthogonal scores and gives the same predictions as using conventional PLS regression, is proposed to be used as a standard to harmonise the interpretation of NIR spectra.

Future research

A theoretical problem to tackle in the future is the curve resolution modelling of NIR spectra from biological material. This analytical tool is of particular interest for example in chemistry where solutions or powders can be mixed in any proportion. But to maintain biological processes in biological materials the mixing, *e.g.* seed-water, can only be done within parts of the theoretical interval. This limits the use of self-modelling curve resolution that aims at producing pure concentrations and pure spectra of components.

For industrial use, conventional NIR spectroscopy is applicable for classification and quantification of seed characteristics, but also high throughput seed sorters that use the NIR region are greatly needed. In production systems of agriculture

and forestry different seed lots usually do not exhibit the same performance in producing germinates, sprouts or plants. This is just like industrial batch processes starting with natural raw materials of varying quality. The supervising systems for monitoring, regulation and control in these industrial processes aim at narrow quality limits in the final product. Multivariate NIR analyses of seeds can support such systems since NIR spectroscopy offers continuous on-line measurements that are fast and non-destructive. The application of NIR techniques together with multivariate statistical process control can further improve seed quality, which saves resources in agriculture and forestry.

Patterns of NIR spectra can be constructed observing time series spectra of seeds in different states of the germination process. These patterns of seed-water interactions describe hydrothermal time dependent processes in seeds. Such spectral patterns can be constructed for viable, non-viable, dry and non-dry seeds within a species, subspecies, variety, ecotype or breeding line as well as for seeds of different geographic origin. By comparing master patterns with spectra of new seeds it will be possible to predict the physiological seed state. It will also give better possibilities to supervise the control and regulation of seed priming processes to reach wanted states before sowing. The concept of NIR pattern recognition can also integrate biochemical analyses, future use of in vivo reporters in seeds, etc.

In order to make new discoveries and to gain new insights of scientific interest the use of hyperspectral NIR imaging gives good opportunities especially if coupled to in vivo reporters and microchemical analyses that can be pixel orientated. Hyperspectral NIR imaging is an advanced tool aimed at fulfilling analyses in a micro-scale targeted to single seed variation for prediction of seed characteristics. Estimations indicate that thousands of genes are involved in the germination processes. If the whole seed cycle from development to germination is included the number of involved genes increases. As gene regulation is differentiated into different seed tissues and cell organelles hyperspectral NIR imaging down to microns in pixel size becomes a useful tool in the research focused on gene mapping and proteomics. It is also shown (Munck *et al*. 2001, Delwiche *et al.* 2002) that the environmental effect in seed phenotypes can be analysed by NIR.

Acknowledgements

I whish to express my sincere gratitude and appreciation to my friend and supervisor associate professor PhD Paul Geladi, recipient of the EAS Chemometrics Award 2002 and one of the 15 most cited scientists in the field of chemometrics, who patiently has introduced me into the multivariate world and shown me how to use these tools in combination with NIR spectroscopy. His never failing encouragement and support has been invaluable in discussions, revisions of manuscripts, solving PLS problems, introducing Matlab, finding funding, meetings with researchers in the fields of chemometrics and NIR spectroscopy and so on. Sincere thanks, Paul.

I greatly appreciate the invaluable support from my friends and co-supervisors Professor Iwan Wästerlund and Professor Björn Hånell at the Department of Silviculture for their wise advices, fruitful discussions and constructive criticism of manuscripts.

I also want to express my gratitude and dedicate special thoughts to:

- Co-author Dr. Riccardo Leardi, Genova, for excellent GA modelling and for giving me a eureka-experience in the selection of wavelengths - I really enjoyed it and his hospitality.
- Co-author Prof. Per-Christer Odén for criticism of manuscripts and part time collaboration in the research project.
- PhD Roger Magnusson for criticism of a manuscript.
- All friends at SLU in Umeå for encouragement, especially the entire staff at the Department for Silviculture for all help during the years:
- Margareta Söderström for excellent technical assistance and cooperation.
- Professor emeritus Per-Ove Bäckström for support in the critical beginning of my NIR studies of seeds and Professor Urban Bergsten for collaboration in developing the PREVAC. My thoughts are also going to the late Professor Milan Simak who introduced me into seed science.
- Lena Walfridsson for technical assistance in collecting data for paper I.
- Karin Strand-Folmerz, Ann-Kathrin Persson and Inga-Lis Johansson for skilful support in administrative matters.
- The staff at the Swedish Forest Research Institute in Sävar for the period when I was responsible for practical seed processing, especially:
- PhD Ola Rosvall who was co-applicant of the research project that financed main parts of my studies; Thyra Mähler and Kjell Andersson for transferring practical and “silent” knowledge of seeds.
- All friends at SLU Omvärld and Uminova for patient understanding of the PhD in progress.
- NIR Nord for seminars, conferences and informal midwinter NIR meetings, especially: Professor Calle Nilsson, FOI, PhD Lars Wallbäcks, Wedometrics, and PhD-student Josefina Nyström, Department of Chemistry, Umeå University, for discussions, support and encouragement.
- Senior researcher Jouko Malinen, VTT Electronics, Oulo, for discussions of NIR instrumentation and PhD Kari Saarinen, ABB Oy Corporate Research, Wasa, for introducing the scattering matrix to me.

- Anne and Jim Burger for linguistic revisions.
- The Reseach Group of Chemometrics, Department of Chemistry at Umeå University for allowing me to use the NIRSystems 6500 spectrometer.
- The staff at the Forest Library and Grafiska Enheten at SLU in Umeå for excellent service.
- SCA Bogrundet Plant Nursery and Svenska Skogsplantor for seed supplies.
- The Swedish Research Council for Environment, Agricultural Research and Spatial Planning, the EU Interreg 3A Kvarken-MittSkandia Unizon-project NIRCE, the Kempe Foundations and Skogssällskapet for funding of the studies.

My mother Eugenia, my late father Verner and my brothers Stig and Robert are acknowledged for invaluable encouragement and support.

Finally, many sincere thanks to my dear wife Åsa and my dear daughters Ylva, Ragna and Hedvig – let's travel.

References

Alberts B., Bray D., Lewis J., Raff M., Roberts, K and Watson, J.D. 1994. *Molecular biology of the cell.* 3rd ed. Garland Publishing Inc., London, UK.

Allen P.S and Meyer S.E. 1998. Ecological aspects of seed dormancy loss. *Seed Science*, **8**, 183-191.

Allen P.S and Meyer S.E. 2002. Ecology and ecological genetics of seed dormancy in downy brome. *Weed Science*, **50**, 241-247.

Alvarado V. and Bradford K.J. 2002. A hydrothermal time model explains the cardinal temperatures for seed germination. *Plant Cell and Environment*, **25**, 1061-1069.

Anonymous. 2000. *User's guide to Simca-P, version 8.0*. Umetrics AB, Umeå, Sweden.

Anonymous. 2002a. Product information. *Elexso Vision Gmb*, Hamburg, Germany.

Anonymous. 2002b. Product information. *Sortex Ltd*, London, UK.

Anonymous 2002c. Seed Scan™. Satake USA Inc., Houston, TX, USA.

Anonymous. 2002d. *User's Guide to SIMCA-P, SIMCA-P+. Version 10.0.* Umetrics AB, Umeå, Sweden.

Anonymous. 2003. *MATLAB 6.5.The Language of Technical Computing.* The MathWorks, Inc., Natick, MA, USA.

Antti H. 1999. *Multivariate characterization of wood related materials*. Dissertation, Department of Chemistry, Umeå University, Umeå, Sweden. ISBN 91-7191-712-8.

Antti H, Sjöström M. and Wallbäcks L. 1996. Multivariate calibration models using NIR pectroscopy on pulp and paper industrial applications. *Journal of Chemometrics*, **10**, 591-603.

von Arnold S., Sabala I., Bozhkov P., Dyachok J. and Filonova L. 2002. Developmental pathways of somatic embryogenesis. *Plant Cell Tissue and Organ Culture*, **69**, 233-249.

Axon T.G, Brown R., Hammond S.V., Maris S.J and Ting F.J. 1998. Focusing near infrared spectroscopy on the business objectives of modern pharmaceutical production. *Journal of Near Infrared Spectroscopy*, **6**, 13–19.

Baker J.E., Dowell F.F. and Throne J.E. 1999. Detection of parasited rice weevils in wheat kernels with near-infrared spectroscopy. *Biological control*, **16**, 88-90.

Barnes B.J., Dhanoa M.S. and Lister S.J. 1989. Standard normal variate transformation and de-trending of near infrared diffuse reflectance spectra. *Applied Spectroscopy*, **43**, 772-777.

Bartley P.G., Nelson S.O., McClendon R.W. and Travelsi, S. 1998. Determining moisture content of wheat with an artificial neural network from microwave transmission measurements. *IEEE Transactions on Instrumentation and Measurement*, **47**, 123-126.

Batlla D. and Benech-Arnold R.L. 2003. A quantitative analysis of dormancy loss dynamics in *Polygonum aviculare* L. seeds: Development of a thermal time model based on changes in seed population thermal parameters. *Seed Science Research*, **13**, 55-68.

Barton F.E., II. 2002. Theory and principles of near infrared spectroscopy. *Spectroscopy Europe*, **14**, 14-18.

Beebe K., Pell R. and Seasholtz M.-B. 1998. *Chemometrics. A Practical Guide*, Wiley, New York.

Benamar A., Tallon C. and Macherel D. 2003. Membrane integrity and oxidative properties of mitochondria isolated from imbibing pea seeds after priming or accelerated ageing. *Seed Science Research*, 13, 35-45.

Ben-Gera I. & Norris K.H. 1968. Determination of moisture in soybeans by direct spectrophotometry. *Israel Journal of Agricultural Research*, **18**, 125-132.

Bergsten, U. 1987. *Incubation of Pinus sylvestris L. and Picea abies L. (Karst) seeds at controlled moisture content as an invigoration step in the IDS method*. Dissertation. Swedish University of Agricultural Sciences, Department of Silviculture, Umeå, Sweden. ISBN 91-576-2989-0.

Bewley J.D. and Black M. 1994. *SEEDS: Physiology of development and germination*. 2nd ed. Plenum Press, New York, USA.

Boelens H.F.M., Kok W.T., de Noord O.E. and Smilde A.K. 2000. Fast on-line analysis of process alkane gas mixtures by NIR spectroscopy. *Applied spectroscopy*, **54**, 406-412.

Bohren C.F. and Huffman D.R. 1998. *Absorption and Scattering of Light by Small Particles*. Wiley Science Paperback Series, New York, USA.

Bokobza, L. 1998. Near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, **6**, 3-17.

Born M. and Wolf E. 1999. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (7thed.). Cambridge University Press, Cambridge, UK.

Bracy R.P. and Parish R.L. 2001. A comparison of seeding uniformity of agronomic and vegetable seeders. *Horttechnology*, **11**, 184-186.¨

Bradford K.J 2002. Applications of hydrothermal time to quantifying and modeling seed germination and dormancy. *Weed Science*, **50**, 248-260.

Bray C.M. 1995. Biochemical processes during the osmopriming of seeds. 1995. In: *Seed Development and Germination.* (Eds: Kigel, J. and Galili, G.). Marcel Dekker Inc., New York, USA. ISBN 0-8247-9229-7.

Brereton R.G.*,* 2003. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, Wiley, Chichester, UK.

Brown P. 1993. *Measurement, Regression and Calibration*. Clarendon Press, Oxford.

Burns D. and Ciurczak E.W. 2001. *Handbook of Near-Infrared Analysis.* 2nd ed. Marcel Dekker, NewYork, USA.

Campbell M.R., Brumm T.J. and Glover D.V. 1997. Whole grain amylose analysis in maize using near-infrared transmittance spectroscopy. *Cereal Chemistry*, **74**, 300-303.

de Castro R.D., van Lammeren A.A.M., Groot S.P.C., Bino R.J. and Hilhorst H.W.M. 2000. Cell division and subsequent radicle protrusion in tomato seeds are inhibited by osmotic stress but DNA synthesis and formation of microtubular cytoskeleton are not. *Plant Physiology*, **122**, 327-335.

Chambers J., Cowe I.A. van Wyk C.B, Wilkin D.R. and Cuthbertson D.C. 1992. Detection of insects in stored products by NIR. In: *Near-Infrared Spectroscopy: Bridging the Gap Between Data Analysis and NIR Applications*. (Eds: Hildrum K.I., Isaksson T., Næs T. and Tandberg A.). Ellis-Horwood, Chichester, UK. 203-208.

Copeland L.O. and McDonald M.B. 2001. *Principles of Seed Science and Technology*. 4th edition. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Curcio J.A. and Petty C.C. 1951. The near infrared absorption spectrum of liquid water. *Journal of Optical Society of America*, **41**, 302-304.

Dåbakk E., Nilsson M., Geladi P., Wold S. and Renberg I. 2000. Inferring lake water chemistry from filtered seston using NIR spectrometry. *Water research*, **34**, 1666-1672.

Davies, A.M.C. 1990. Subdivisions of the infrared region. *Applied Spectroscopy*, **44**, A14.

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

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

Delwiche S.R., Graybosch R.A. and Peterson C.J. 1999. Identification of wheat lines possessing the 1AL.1RS or 1BL.1RS wheat-rye translocation by near-infrared. *Cereal Chemistry*, **76**, 255-260.

Delwiche S.R., Graybosch R.A., Nelson L.A. and Hruschka W.R. 2002. Environmental effects on developing wheat as sensed by near-infrared reflectance of mature grains. *Cereal Chemistry*, **79**, 885-891.

Diamantaras K.I. and Kung, S.Y. 1996. *Principal component neural networks: theory and applications*. Wiley & Sons, Inc. N.Y., USA. ISBN 0-471-05436-4.

Dowell F.E, 1998. Automated colour classification of single wheat kernels using visible and near-infrared reflectance. *Cereal Chemistry*, **75,** 142-144.

Dowell F.E., Throne J.E. and Baker, J.E. 1998. Automated non-destructive detection of internal insect infestation of wheat kernels by using near-infrared reflectance spectroscopy. *Journal of Economic Entomology*, **91**, 899-904.

Dowell, F.E., Throne, J.E. and Baker, J.E. 1999. Identifying stored-grain insects using near-infrared spectroscopy. *Journal of Economic Entomology*, **92**, 165-169.

Dowell FE, Pearson TC, Maghirang EB, Xie F, Wicklow DT. 2002. Reflectance and transmittance spectroscopy applied to detecting fumonisin in single corn kernels infected with Fusarium verticillioides. *Cereal Chemistry*, **79**, 222-226.

Downey G. 1985. Estimation of moisture in undried wheat and barley by near infrared reflectance. *Journal of the Science of Food and Agriculture* **36**, 951-958.

Downey G. 1996. Non-invasive and non-destructive percutaneous analysis of farmed salmon flesh by near infra-red spectroscopy. *Food Chemistry*, **55**, 305-311.

Downey G., Robert P., Bertrand D. and Kelly P.M. 1990. Classification of commercial skim milk powders according to heat-treatment using factorial discriminant analysis of near-infrared reflectance spectra. *Applied Spectroscopy*, **44**, 150-155.

Downie B., Coleman J., Scheer G., Wang B.S.P., Jensen M. and Dhir N. 1998. Alleviation of seed dormancy in white spruce (*Picea glauca* [Moench.] Voss.) is dependent on the degree of seed hydration. *Seed Science and Technology,* **26**, 555-569.

Draper N. and Smith H. 1981. *Applied Regression Analysis* (2nd ed). Wiley, New York.

Efimov, Y.Y. 2001. Asymmetry in H_2O molecules in the liquid state and its consequences. *Journal of Structural Chemistry*, **42**, 935-945.

Efimov Y.Y. and Naberukhin Y.I. 2002. On the interrelation between frequencies of stretching and bending vibrations in liquid water. *Spectrochimica Acta Part A – Molecular and Biomolecular Spectroscopy*, **58**, 519-524.

Eriksson L., Johansson E., Kettaneh-Wold K. and Wold S. 1999. *Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA and PLS)*. Umetrics AB, Umeå, Sweden.

Ergon R. 2002. PLS score-loading correspondence and bi-orthogonal factorization. *Journal of Chemometrics* **16**, 368-373.

Esbensen K.H. 2000. *Multivariate Data Analysis in Practice: An Introduction to Multivariate Data Analysis and Experimental Design*. 4th edition, CAMO ASA, Norway. 598 pp.

Eskeröd 1973. Jordbruk under femtusen år, redskapen och maskinerna. ([Swedish] Agriculture over five thousand years, the equipment and the machines.) LTs Förlag, Borås, Sweden.

Espinosa A., Sanchez M., Osta S., Boniface C., Gil J., Martens A., Descales B., Lambert D. and Valleur M. 1994. On-line NIR analysis and advanced control improve gasoline blending. Oil & Gas Journal, **92**, 49-56.

Falk K.L., Behal R.H., Xiang C.B. and Oliver D.J. 1998. Metabolic bypass of the tricarboxylic acid cycle during lipid mobilization in germinating oilseeds - Regulation of NAD(+)-dependent isocitrate dehydrogenase versus fumarase. *Plant Physiology*, **117**, 473-481.

Falleri E. and Pacella R. 1997. Applying the IDS method for the removal of empty seeds in Platanus X acerifolia. *Canadian Journal of Forest Research*, **27**, 1311-1315.

FAO (Food and Agricultural Organization). 2001. Global Forest Resources Assessment: 2000 main report. Rome, Italy.

FAOSAT. 2003. On-Line Statistical Database *http://apps.fao.org*. Food and Agricultural Organization, Rome, Italy (2003-08-05).

Fearn T. 2000. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, **50**, 47-52.

Finch-Savage W.E., Steckel J.R.A. and Phelps L. 1998. Germination and post-germination growth to carrot seedling emergence: predictive threshold models and sources of variation between sowing occasions. *New Phytologist* **139**, 505-516.

Finch-Savage W.E., Phelps L., Peach L. & Steckel J.R.A. 2000. Use of threshold germination models under variable field conditions. In: *Seed Biology: Advances and Applications.* (Eds: Black M., Bradford K.J and Vázquez-Ramos). CAB International, Wallingford, Oxon, UK.

Font R., Rio M.D., Fernandez-Martinez J.M., Haro A.D., del Rio M. and De Haro A. 1999. Using near infrared spectroscopy (NIRS) for the determination of bulk density in Indian mustard seed. *Cruciferae Newsletter*, **21**, 75-76.

Forrest S. 1993. Genetical algorithms – principles of natural selection applied to computation. *Science* **261**, 872-878.

Frank I.E. 1987. Intermediate least-squares regression metod. *Chemometrics and Intelligent Laboratory Systems*, **1**, 233-242.

Geladi P. 1988. Notes on the history and nature of partial least squares (PLS) modelling. *Journal of Chemometrics*, **2**, 231-246.

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

Geladi P., Bärring H., Dåbakk E., Trygg J., Antti H., Wold S., Karlberg B. 1999. Calibration transfer for predicting lake-water pH from near infrared spectra of lake sediments. *Journal of Near Infrared Spectroscopy*, **7**, 251-264.

Geladi P., Nyström J., Eriksson J., Nilsson A., Lithner F. and Lindholm-Sethson B. 2000. A first multivariate NIR study of skin condition in diabetic patients and controls. *Journal of Near Infrared Spectroscopy*, **8**, 217-227.

Ghaedian A.R. and Wehling R.L. 1997 Discrimination of sound and granary-weevil-larva-infested wheat kernels by near-infrared diffuse reflectance spectroscopy. *Journal of AOAC International*, **80**, 997-1005.

Granström, A. and Schimmel, J. 1993. Heat-effects on seed and rhizomes of a selection of boreal forest plants and potential reaction to fire. *Oecologia*, **94**, 307-313.

Gummerson, R.J. 1986. The effect of constant temperatures and osmotic potentials on the germination of sugar beet. *Journal of Experimental Botany* **37**, 729-741.

Hagner, M. 1981. *The use of germinated seeds increases nursery efficiency.* Report No. 121, Department of Forest Science, Umeå University, Umeå, Sweden.

Halsey S.A. 1987. Analysis of whole barley kernels using near-infrared reflectance spectroscopy. *Journal of the Institute of Brewing*, **93**, 416-464.

Hampton J.G. and TeKrony D.G. 1995. *Handbook of Vigour Test Methods*. International Seed Testing Association (ISTA), Zurich, Switzerland.

Hannerz M., Eriksson U., Wennström U. and Wilhelmsson L. 2000. *Scots pine and Norway spruce seed orchards in Sweden – a description with an analysis of future seed supply*. Report 1, The Forestry Research Institute of Sweden, Uppsala, Sweden.

Hardegree S.P. and Emmerich W.E. 1990. Effect of polyethylene-glycol exclusion on the water potential of solution-saturated filter-paper. *Plant Physiology*, **92**, 462-466.

Harmond J.E., Brandenburgh N.R. and Klein L.M. 1968. *Mechanical seed cleaning and handling*. Agriculture Handbook. No. 354. USDA, ARC, Washington DC. USA.

Hazen K.H., Arnold M.A., Small G.W. 1998. Measurement of glucose and other analytes in undiluted human serum with near-infrared transmission spectroscopy *Analytica Chimica Acta*, **371**, 255-267.

Heise H.M., Marbach R. and Bittner A. 1998. Clinical chemistry and near infrared spectroscopy: multicomponent assay for human plasma and its evaluation for the determination of blood substrates. *Journal of Near Infrared Spectroscopy*, **6**, 361–374.

Heydecker W., Higgings J. and Turner, Y.J. 1975. Invigoration of seeds. *Seed Science and Technology*, **3**, 881-888.

Hindle, P.H. 2001. Historical development. In**:** *Handbook of Near-infrared Spectroscopy* (Eds. Burns D.A. and Ciurcazk E.W.). 2nd ed. Mercel Dekker Inc., New York. p 1-6.

Hindle P. H. 2002. The last millennium: a brief history of science leading to current infrared technology. *Near Infrared Spectroscopy: Proceedings of the 9th International*

Conference. (Eds: Davies A.M.C. and Giangiacomo R.). NIR Publications, Chichester, West Sussex, UK.

Hirano S., Okawara N., Narazaki S. 1998. Near infrared detection of internally moldy nuts. *Bioscience, Biotechnology and Biochemistry*, **62**, 102-107.

Hopkins D. 2001. Derivatives in spectroscopy. *Near Infrared Spectroscopy*, **2**, 1-13.

Höskuldsson A. 1996. *Prediction Methods in Science and Technology. Vol.1 - Basic Theory*. Thor Publishing, Copenhagen, Denmark.

Hühn M. 2001. Effects of nonregular spatial distribution of plants on yield per area: A theoretical approach with applications to winter oilseed rape (*Brassica napus* L.). *Journal of Agronomy and Crop Science*, **187**, 241-250.

Hull E.L., Conover D.L. and Foster T.H. 1999. Carbogen-induced changes in rat mammary tumour oxygenation reported by near infrared spectroscopy. *Brittish Journal of Cancer*, **79**, 1709-1716.

Hurburgh C.R. Jr., Wu Y. and Siska, J. 1995. Effect of seed size and density on near-infrared transmittance analysis of corn and soybeans. *Applied Engineering in Agriculture*, **11**, 677-684.

ISTA (International Seed Testing Association). 1999. International rules for seed testing. *Seed Science and Technology*, **27**, Supplement.

Iyer M., Morris H.R. and Drennen J.K. 2002. Solid dosage form analysis by near infrared spectroscopy: comparison of reflectance and transmittance measurements including the determination of effective sample mass. *Journal of Near Infrared Spectroscopy*, **10** (4): 233-245.

Jalink H., van der Schoor R., Frandas A., van Pijlen J.G., Bino R.J. 1998. Chlorophyll fluorescence of Brassica oleracea seeds as a non-destructive marker for seed maturity and seed performance. *Seed Science Research*, **8**, 437-443.

Jedvert I., Josefson M. and Langkilde F.J. 1998. Quantification of an active substance in a tablet by NIR and Raman spectroscopy. *Journal of Near Infrared Spectroscopy*, **6**, 279–289.

Jolliffe I.T. 1986. *Principal Component Analysis*. Springer Verlag, Berlin, Germany.

Jöreskog K. and H. Wold. 1982. Systems Under Indirect Observation. Causality, Structure, Prediction, Part II. North-Holland, Amsterdam.

Kachman S.D. and Smith J.A. 1995. Alternative measures of accuracy in plant spacing for planters using single seed metering. *Transactions of the ASAE*, **38**, 379-387.

Kawamura S., Natsuga M. and Itoh K. 1998. Visual and near-infrared reflectance spectroscopy for determining physiochemical properties of rice. *American Society of Agricultural Engineers (ASAE) Annual Meeting, Orlando, Florida, USA. 12-16 July, 1998*. ASAE Paper no. 983063. 9pp.

Kawano S.2002. Application to agricultural products and foodstuffs. In: *Near Infrared Spectroscopy, Principles, Instruments, Application*. (Eds: Siesler H., Ozaki Y., Kawata S. and Heise H.). Wiley-VCH, Chichester, UK. p 269-287.

Kigel, J. and Galili, G. 1995. *Seed development and germination.* Marcel Dekker Inc., New York, USA. ISBN 0-8247-9229-7.

Kim J.G., Zhao D.W., Song Y.L., Constantinescu A., Mason R.P. and Liu H.L. 2003. Interplay of tumor vascular oxygenation and tumor pO(2) observed using near-infrared spectroscopy, an oxygen needle electrode, and F-19 MR pO(2) mapping. *Journal of Biomedical Optics*, **8**, 53-62.

King R.J., King K.V. and Woo K. 1992. Microwave moisture measurement of grains. *IEEE Transactions on Instrumentation and Measurement*, **41**, 111-115.

Kohel, R.J. 1998. Evaluation of near-infrared reflectance for oil content of cottonseed. *Journal of Cotton Science*, **2**, 23-26.

Konstantinova P., Van der Schoor R., Van den Bulk R. and Jalink H. 2002. Chlorophyll fluorescence sorting as a method for improvement of barley (*Hordeum vulgare* L.) seed health and germination. *Seed Science and Technology*, **30**, 411-421.

Kwon Y.K. and Cho R.K. 1998. Identification of rice variety using near infrared spectroscopy. In: *Proceedings of NIR-97*. (Ed: Davies A.M.C.) *Journal of Near Infrared Spectroscopy*, **6**, A67-A73.

Lamb D.T. and Hurburgh C.R. 1991. Moisture determination in single soybean seeds by near infrared transmittance. *Transactions of the ASAE*, **34** (5), 2123-2129.

Lawrence K.C., Windham W.R., Nelson S.O. 1998a. Wheat moisture determination by 1- to 110-MHz swept-frequency admittance measurements. *Transactions of the ASAE,* **41**, 135-142.

Lawrence K.C., Nelson S.O., Bartley P.G. 1998b. Measuring dielectric properties of hard red winter wheat from 1 to 350 MHz with a flow-through coaxial sample holder. *Transactions of the ASAE*, **41**, 143-150.

Leardi R. 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, **14**, 643-655.

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

Leardi R. and González A.L. 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems*, **41**, 195-207.

Leardi R. and Nørgaard L. 2003. *Sequential application of backward interval-PLS and Genetic Algorithms for wavelength selection.* Colloquium Chemiometricum Mediterraneum V, June 25-27, 2003, Book of Abstracts, O2.

Lestander, T. 1986. Applications of new methods for seed treatment in order to exploit the most suitable Scots pine provenances. *Proceedings of the Frans Kempe Symposia, Umeå. June 10-11 1986.* Report 6, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Umeå, Sweden. p 179-188.

Lestander, T. 1988. *Utveckling av utrustning för att i stor skala avlägsna dött respektive skadat frö samt höja fröets groningshastighet*. ([Swedish] Development of equipment for large scale removal of dead and damaged seed, respectively, and to improve the germination speed of seeds). Internrapport 198, Institutet för skogsförbättring (Institute of Forest Tree Improvement), Sävar, Sweden. 26pp.

Lestander T.A., Geladi P. and Odén P.C. 2002. 2- and 3- way analysis of NIR scans from seed crossings. *Proceedings of the 10th International Conference on Near-Infrared Spectroscopy*. (Eds: Davies A.M.C. and Cho R.K., NIR Publications). Chichester, West Sussex, UK. p. 385-388.

Li B.B., Morris A.J. and Martin E.B. 2002. Orthogonal signal correction: algorithmic aspects and properties. *Journal of Chemometrics*, **16**, 556-561.

Lindgren F., Geladi P., Rännar S. and Wold S. 1994. Interactive variable selection (IVS) for PLS. Part I: Theory and algorithms. *Journal of Chemometrics* **8**, 349-363.

Lindgren F., Geladi P., Berglund A., Sjöström M. and Wold S. 1995. Interactive variable selection (IVS) for PLS. Part II: chemical applications. *Journal of Chemometrics* **9**, 331-342.

Logan D.C., Millar A.H., Sweetlove L.J., Hill S.A. and Leaver C.J. 2001. Mitochondrial biogenesis during germination in maize embryos. *Plant Physiology*, **125**, 662-672.

Malabadi R.B. and Nataraja K. 2002. Large scale production and storability of encapsulated somatic embryos of mothbean (*Vigna aconitifolia* Jacq). *Journal of Plant Biochemistry and Biotechnology*, **11**, 61-64.

Marbach, R. 1993. Messverfahren zur IR-spektroskopischen Blutglucosebestimmung. ([German] A measurement method for determination of blood glucose by IR spectroscopy). *Fortschr.-Ber. VDI Verag, Series 8, Vol. 346*, Düsseldorf, Germany. ISBN 3-18-144608-4.

Martens H. and S.-Å. Jensen.1983. Partial Least Squares regression: a new two-stage NIR calibration method. In: *Progress in Cereal Chemistry and Technology*. Eds. Holas J. and Kratochvil J. Elsevier, Amsterdam. p. 607-647.

Martens H. and Næs T. 1987. Multivariate calibration by data compression. In: *Near-infrared technology in agricultural and food industries.* (Ed: Williams P.C. and Norris K.) American Association of Cereal Chemists, St. Paul, Minnesota, USA. p 57-87.

Martens H. and Næs T. 1989. *Multivariate calibration.* John Wiley & Sons Ltd, Chichester, UK.

Martens H. and Martens M. 2000. *Multivariate Analysis of Quality, An Introduction*. Wiley, Chichester, UK.

McCaig T.N. 2002. Extending the use of visible/near-infrared reflectance spectrophotometers to measure colour of food and agricultural products. *Food Research International*, **35,** 731-736.

McClure W.F. 1994. Near-infrared spectroscopy – the giant is running strong. *Analytical Chemistry*, **66**, A43-A53.

McDonald M.B., Sullivan J. and Laurer, M.J. 1994. The pathway of water-uptake in maize seeds. *Seed Science and Technology*, **22**, 79-90 1994.

McDonald, M.B. 1999. Seed deterioration: physiology, repair and assessment. *Seed Science and Technology*, **27** (1): 177-237.

Michel B.E. 1983. Evaluation of the water potentials of solutions of polyethylene glycol 8000 both in absence and presence of other solutes. *Plant Physiology*, **72**, 66-70.

Mueller H. 1948. The foundation of optics. *J. Opt. Soc. Am*. **38**, 661.

Munck L., Pram Nielsen J., Møller B., Jacobsen S., Søndergaard I., Engelsen S. B., Nørgaard L. and Bro R. 2001. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Analytica Chimica Acta*, **446**, 171-186.

Nelson S.O. and Lawrence K.C. 1994. RF impedance and DC conductance determination of moisture in individual soybeans. *Transactions of the ASAE*, **37**, 79-182.

Nilsson M.B., Dåbakk E., Korsman T. and Renberg I. 1996. Quantifying relationships between near-infrared reflectance spectra of lake sediments and water chemistry. *Environmental Science & Technology*, **30**, 2586-2590.

Nord S.P., DuBreuil R., Anna G. Cavinato A.G., Mayes D.M., Lin M. and Rasco B. 2002. Penetration depth studies in cod tissue using shortwave near infrared spectroscopy. *Eastern Oregon Science Journal*, **17**, 37-41.

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

Norris. K.H. and Butler W.L. 1961. Techniques for obtaining absorbance spectra on intact biological samples. *IRE Transactions on Bio-Medical Electronics*, **8**, 153-157.

Norris K.H. and Hart J.R. 1996. Direct spectrophotometric determination of moisture content of grain and seeds. *Journal of Near Infrared Spectroscopy*, **4**, 23-30. (reprint from 1965 in *Principles and Methods of Measuring Moisture Content in Liquds and Solids* (Ed: Waxler, A.) Vol IV. Reinhold Publishing Corporation, New York, USA. p 19-25).

Norris K.H. 1988. History, present status, and future prospects for NIRS. In: *Analytical Applications of Spectroscopy*. (Eds: Creaser C.S. and Davies A.M.C.). Royal Society of Chemistry, London, UK. p 3-8.

Nyström C. 1982. Biplantor - en fara för stabilitet och rotutveckling? ([Swedish] Multiple seeding in the nursery - a threat for future tree stability and root growth.) *Plantnytt*, **2**, 1-4.

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

Olinger J.M., Griffiths, P.R and Burger, T. 2001. Theory of diffuse reflection in the NIR region. In *Handbook of Near-Infrared Spectroscopy* (eds. D.A. Burns and E.W. Ciurczak). pp. 19-52. Marcel Dekker Inc., New York.

Öquist G, and Wass R. 1988. A portable, microprocessor operated instrument for measuring chlorophyll fluorescence kinetics in stress physiology. *Physiologia Plantarum*, **73**, 211-217.

Osborne B.G., Fearn T. and Hindle, P.H. 1993. *Practical NIR spectroscopy with applications in food and beverage analysis*. 2nd ed. Longman Scientific & Technical, Harlow, Essex, England.

Ozmerzi A., Karayel D. and Topakci M. 2002. Effect of sowing depth on precision seeder uniformity. *Biosystem Engineering*, **82**, 227-230.

Pasikatan M.C. and Dowell F.E. 2001. Sorting systems based on optical methods for detecting and removing seeds infested internally by insects or fungi: A review. *Applied Spectroscopy Reviews*, **36**, 399-416.

Pazdernik D.L., Killam A.S. and Orf J.H. 1997. Analysis of amino and fatty acid composition in soybean seed, using near infrared reflectance spectroscopy. *Agronomy Journal*, **89**, 679-685.

Pearson, T.C. 1999. Use of near-infrared transmittance to automatically detect almonds with concealed damage. *Lebensmittel - Wissenschaft und Technologie*, **32**, 73-78.

Pearson T.C., Wicklow D.T., Maghirang E.B., Xie F., Dowell F.E. 2001. Detecting aflatoxin in single corn kernels by transmittance and reflectance spectroscopy. *Transactions of the ASAE*, **44**, 1247-1254.

Powers J.B., Gunn J.T. and Jacob F.C. 1953. Electronic colour sorting of fruits and vegetables. *Agricultural Engineering*, **34**, 149-154, 158.

Reich G. 2002. Potential of attenuated total reflection infrared and near-infrared spectroscopic imaging for quality assurance/quality control of solid pharmaceutical dosage forms. *Pharmazeutische Industrie*, **64**, 870-874.

Reich P.B., Oleksyn J. and Tjoelker M.G. 1994. Seed mass effects on germination and growth of diverse European Scots pine populations. *Canadian Journal of Forest Research*, **24**, 306-320.

Repo T., Paine D.H. and Taylor A.G. 2002. Electrical impedance spectroscopy in relation to seed viability and moisture content in snap bean (*Phaseolus vulgaris* L.). *Seed Science Research*, **12**, 17-29.

Ridgway C., Chambers J. and Cowe I.A. 1999. Detection of grain weevils inside single wheat kernels by a very near infrared two-wavelength model. *Journal of Near Infrared Spectroscopy*, **7**, 213 -221.

Rowse H.R and Finch-Savage W.E. 2003. Hydrothermal threshold models can describe the germination response of carrot (*Daucus carota*) and onion (*Allium cepa*) seed populations across both sub- and supra-optimal temperatures. *New Phytologist*, **158**, 101-108.

Rumler C., Gregorova E., Turzik Z., Resatko M. and Mazanek M. 1993. Multiradiometric measurement of selected characteristics of forest tree seeds. *Lesnictvi*, **39**, 217-221.

Salter P.J. 1978. Fluid drilling of pre-germinated seeds: process and possibilities. *Acta Horticulturae*, **83**, 245-250.

Sato T. 1994. Application of principal-component analysis on near-infrared spectroscopic data of vegetable-oils for their classification. *Journal of the American Oil Chemists Society*, **71**, 293-298.

Sato T., Uezono I., Morishita T. and Tetsuka T. 1998. Non-destructive estimation of fatty acid composition in seeds of *Brassica napus* L. by near-infrared spectroscopy. *Journal of the American Oil Chemist´s Society,* **75**, 1877-1881.

Segtnan V.H., Sasic S., Isaksson T. and Ozaki Y. 2001. Studies on the structure of water using two-dimensional near-infrared correlation spectroscopy and principal component analysis. *Analytical Chemistry*, **73**, 3153-3161.

Shen, T.Y. 2000. *Forest Tree Seed Quality Determination Based on Enzyme Activities*. Doctoral Thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden. ISBN 91-576-5891-9.

Shenk J.S., Workman J.J. and Weterhaus M.O. 2001. Application of NIR spectroscopy to agricultural products. In: *Handbook of Near-Infrared Analysis.* (Eds: Burns D. and Ciurczak E.W.). 2nd ed. Marcel Dekker, NewYork., USA. 419-474.

Simak, M. 1981. Bortsortering av matat-dött frö ur ett fröparti. (Removal of filled-dead seeds from a seed bulk.) *Sveriges Skogsvårdsförbunds Tidskrift*, **5**, 31-36.

Simak, M. 1984. A method for removal of filled-dead seeds from a sample of *Pinus contorta*. *Seed Science and Technology*, **12**, 767-775.

Simak M., Bergsten U. and Henriksson G. 1989. Evaluation of ungerminated seeds at the end of germination test by radiography. *Seed science and Technology*, **17**, 361-369.

Sjöblom J., Svensson O., Josefson M., Kullberg H. and Wold S. 1998. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems* **44**, 229-244.

Sjöström, M., Wold, S. and Sjöström, B. 1986. PLS discriminant plots. In: *Pattern Recognition in Practice II.* Elsevier Science Publisher B.V., Holland.

Soltani, A. 2003. *Improvement of Seed Germination of Fagus orientalis Lipsky*. Doctoral Thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden. ISBN 91-576-6509-5.

Stokes, G.G. 1852. On the composition and resolution of streams of polarized light from different sources. *Trans. Camb. Philos. Soc*., **9**, 399-416 (reprinted in *Mathematical and Physical Papers*, Vol. III, Cambridge University Press, Cambridge, UK. 1901).

Sundblad L.G., Sjöström M., Malmberg G. and Öquist, G. 1990. Prediction of frost hardiness in seedlings of Scots pine (*Pinus sylvestris*) using multivariate analysis of chlorophyll-A fluorescence and luminescence kinetics. *Canadian Journal of Forest Research*, **20**, 592-597.

Svensson O., Kourti T., MacGregor J.F. 2002. An investigation of orthogonal signal correction algorithms and their characteristics. *Journal of Chemometrics*, **16**, 176-188.

Szmidt A.E., Aldén T. & Hällgren J.-E. 1987. Paternal inheritance of chloroplast DNA in *Larix*. *Plant Molecular Biology*, **9**, 59-64.

Tang H.W., Ye Y., Li T., Zhou J.S. and Chen G.Q. 2003. Study on Schiff base complexes-cellular DNA interactions by a novel system of Hadamard transform fluorescence image microscopy. *Analyst*, **128**, 974-979.

Taylor, A.G., Churchill, D.B., Lee, S.S., Bilsland, D.M. and Copper, T.M. 1993. Color sorting of coated *Brassica* seeds by fluorescent sinapine leakage to improve germination. *Journal of the American Society for Horticultural Science*, **118**, 551-556.

Taylor A.G., Allen P.S., Bennett M.A., Bradford K.J., Burris J.S. and Misra M.K. 1998. Seed enhancements. *Seed Science Research*, **8**, 245-256.

Thomas T.H. 1983. Stimulation of celeriac and celery seed germination by growth regulator seed soaks. *Seed Science and Technology*, **11**, 301-305.

Thygesen L.G. 1994. Determination of dry matter content and basic density of Norway spruce by near infrared reflectance and transmittance spectroscopy. *Journal of Near Infrared Spectroscopy*, **2**, 127–135.

Tigabu, M. 2003. *Characterization of Forest Tree Seed Quality with Near Infrared Spectroscopy and Multivariate Analysis*. Doctoral Thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Tigabu M. and Odén P.C. 2002. Multivariate classification of sound and insect-infested seeds of a tropical multipurpose tree, *Cordia africana*, with near infrared reflectance spectroscopy. *Journal of Near Infrared Spectroscopy*, **10**, 45-51.

Tillman-Sutela E. and Kauppi A. 1995. The morphological background to imbibition in seeds of *Pinus sylvestris* L. of different provenances. *Trees*, **9**, 123-133.

Tillman-Sutela E. , Johansson A., Laakso P. Mattila T. and Kallio H. 1995. Triacylglycerols in the seeds of northern Scots pine, *Pinus sylvestris* L., and Norway spruce, *Picea abies* (L.) Karst. *Trees*, **10**, 40-45.

Trygg, J. 2001. *Parsimonious Multivariate Models*. Dissertation, Department of Chemistry, Umeå University, Umeå, Sweden. ISBN 91-7305-082-2.

Turza S., Tóth ÁI. and Váradi M. 1998. Multivariate classification of different soyabean varieties. In: *Proceedings of NIR-97*. (Ed: Davies A.M.C.) *Journal of Near Infrared Spectroscopy*, **6**, 183-187.

Urmstrom, D. 1997. A perspective on the direction of the American Seed Industry. *Introduction to the Symposium on Seed Biology and Technology. Fort Collins, Colorado, August, 1997.* American Seed Trade Association, Washington D.C., USA.

Velasco L., Matthaus B. and Mollers, C. 1998. Non-destructive assessment of sinapic acid esters in Brassica species. I. Analysis by near infrared reflectance spectroscopy. *Crop Science*, **38**, 1645-1650.

Velasco L., Mollers C and Becker H.C. 1999. Estimation of seed weight, oil content and fatty acid composition in intact single seeds of rapeseed (*Brassica napus* L.) by near-infrared reflectance spectroscopy. *Euphytica*, **106**, 79-85.

Wallbäcks L., Edlund U., Norden B. and Berglund I. 1991. Multivariate characterization of pulp using solid-state ^{13}C-MNR, FTIR and NIR. *Tappi Journal*, **74**, 201-206.

Welbaum G.E., Bradford K.J., Yim K., Booth D.T. and Oluoch M. 1998. Biophysical, physiological and biochemical processes regulating seed germination. *Seed Science*, **8**, 161-172.

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

Williams P.C., Norris K.H. and Zarowski W.S. 1982. Influence of temperature on estimation of protein and moisture in wheat. *Cereal Chemistry*, **59**:473-477.

Williams P.C., Norris, K.H. and Sobering D.C. 1985. Determination of protein and moisture in wheat and barley by near-infrared transmission. *Journal of Agriculture and Food Chemistry*, 33, 239-244.

William P.C. and Norris K.H. 1987. *Near-Infrared Technology in the Agricultural and Food Industries.* American Association of Cereal Chemists Inc., St. Paul, MN, USA.

Wise B.M. and Gallagher N.B. 1998. *PLS_Toolbox Verson 2.0*. Eigenvector Research Inc., Manson, WA, USA.

Wold H. 1975. Path models with latent variables: the NIPALS approach. In: *Quantitative Sociology*. (Eds: Blalock H., Abenagian A., Borodkin F., Boudon R. and Capecchi V.). Academic Press, New York, USA. p 305-357.

Wold S., Martens, H. and Wold H. 1983. The multivariate calibration problem in chemistry solved by the PLS method. *Proc. Conf. Matrix pencils March 1982.* (Eds: Ruhe A. and Kågström B.). Lecture Notes in Mathematics, Springer Verlag, Heidelberg, p 286-293.

Wold S., Antti H., Lindgren F. and Öhman J. 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **44**, 175-186.

Workman J.J. and Burns D.A. 2001. Commercial NIR instrumentation. In: *Handbook of Near-Infrared Analysis.* 2nd (Eds: Burns D. and Ciurczak E.W.). Marcel Dekker, NewYork., USA. p 53-70.

Wu L.G., Hallgren S.W., Ferris D.M. and Conway K.E. 2001. Effects of moist chilling and solid matrix priming on germination of loblolly pine (*Pinus taeda* L.) seeds. *New Forests*, **21**, 1-16.