Using Hyperspectral imaging for characterisation of wheat kernels.

MASTER'S THESIS

Written by: Frank Manfred Hansper
Supervisor: Associate Professor Sergey Kucheryavskiy

Submitted: 07/06-2019

AALBORG UNIVERSITET

Abstract

The goal of this master thesis was to develop a mathematical model that based on hyperspectral imaging in the near infrared region could characterize wheat kernels from ten different sorts for three chosen quality parameters: relative hardness, protein content and vitreousness. The technology of hyperspectral imaging combines the best of imaging and spectroscopic technologies. It enables the prediction of quality parameters on individual objects internally and non-destructively. The vision for this technology is to use it as a new state-of-the-art method for in-line rapid separation and quality control of products from the agricultural industries. Therefore, it is of upmost importance to test the prediction ability of this technology.

Wheat was chosen as the object of interest because of the variation from grain to grain and an interest from the project partner CSORT to explore the possibilities of this technology on Russian wheat. The wheat was provided by CSORT together with additional parameters of them such as hardness class and Kjeldahl protein. The parameters were chosen based on available literature, their importance in grading and the possibility to measure all three parameters on the same kernels. Though preliminary experiments the reference methods for the parameter's protein content, relative hardness and vitreousness were chosen as follows:

- Protein content: Flow injection analysis.
  This reference method was chosen because of stability and timesaving compared to other available methods. The preliminary experiments showcased a systematic underprediction to Kjeldahl.
- Relative hardness: Rupture force measurement by simple compression.
  This reference method was chosen because of being the only method to measure relative hardness on wheat kernels without being fully destructive. Preliminary experiments showcased that many external factors influenced the measurements such as weight and orientation.
- Vitreousness: Visual evaluation.
  Approved as a standard method. Specialised equipment was not available to measure this parameter properly.

150 kernels were randomly and equally chosen from the ten sorts. 100 were to be used for model calibration and 50 as a test set. Under Image acquisition 142 wavelengths in the NIR of 938–1600 were available with spatial resolution of 2251x320. Four orientations were chosen.

The main experiments showcased that there was good prediction regarding protein content giving models with R^2 for cross-validation as high as 0.775. The relative hardness prediction models were completely random and could not predict this parameter. It is suggested that the external factors critically influenced this parameter making the reference method unsuitable for relative hardness prediction. Regarding virtuousness, after relative hardness measurements many of the kernels were too damaged to visually estimate the vitreousness class. This gave okay to bad prediction models with a misclassification between non-vitreous and fully vitreous on 16.2 % for the test set. However, classification based on wheat hardness class gave 100% correct classification.

Literature has showcased that better predictions on hardness and vitreousness can be made with other reference methods. Therefore, it is suggested to predict two quality parameters instead with proper reference methods for better prediction.

Preface

I would like to give a special thanks to my supervisor Sergey V. Kucheryavskiy for his many helpful discussions and inputs to form this master thesis. I would also like to give a special thanks to professor Jose Manuel Amigo Rubio for using his hyperspectral scanner in Copenhagen and the following explanation of some theoretical aspects. I would also like to give thanks to all laboratory staff at Aalborg University Esbjerg for their great help in giving proper introduction to all the diverse laboratory equipment. I would like to give my thanks to Sergei Zhilin and the company CSORT for providing the sample wheat for the experiments and the opportunity to work with this interesting problem. Lastly, I would like to give my thanks to Novozymes for grating me a scholarship and thereby supporting this tropic of research.

If there is special interest to refabricate the results present in this master thesis, please feel free to contact me on my mail: [redacted]. I will then try to provide the hypercube data and assist in any questions you might have. Because of practicalities and the size of the numerical data (~12Gb), the hypercubes have not been attached to the thesis. They will be stored an unspecified amount of years after the hand-in of 7. June 2019.

In this master thesis, there has been done a great effort to use mainly illustrations under the creative common or public copyright licenses. However, there are used illustrations that fall under copyright law. These illustrations are used solely for research and private studies only to explain a theoretical subject to readers by commenting on the illustrations. All illustrations will be proper credited. If this doesn’t fall under fair use, I will apologize for any copyright infringement and take the necessary actions to uphold the present copyright law. If a permission has been given to use an image, this will be stated in the figure text. Permissions to images are included at the end of this thesis after the appendix chapter.

Table of Contents

1 Introduction

Since the dawn of man, there has been an ever-increasing need for food for man to flourish. It is the food and agricultural industries that has kept putting food on the tables of many. It has always been a battle for these industries to not just ensure there is enough food, but also that the food is high quality, the production is consistent and, most importantly, that the food is safe to eat. All these conditions must be meet with a sustainable production. There is need for technology that helps classify and determine properties of food in a non-destructive rapid manner to get consistency and high quality on an enormous scale. Nowadays imaging and spectroscopy technologies have been used with success to tackle this problem. Imaging used for classification and spectroscopy used for bulk property determination. However, the conventional technologies have been used separately and both have a weak point the other technology excels at. So, has the time come for these technologies to unite and live in symbiosis to create a better day of tomorrow? It is this question that this master thesis will try to answer.

Imaging and spectroscopic methods have shown a big advancement with regard to the agricultural/food industry in recent years. A rapid, real-time online analysis using near-infrared spectrometry is already established for the detection of the chemical and physical properties in different fruits [1]. Areas for the use of imaging technology include monitoring, inspection and grading (based on colour, size, shape or surface texture). This is done by using images to quantify/classify objects by using the spatial properties of them. However, classic image technology without spectrometry have a hard time detecting chemical or physical properties for different food products such as protein content, hardness, firmness, moisture, sugar or acid. Spectrometry on the other hand, have a hard time analysing the internal variation of a food products [2]. When combining these two technologies a state-of-the-art technology arises called hyper spectral imaging (HSI). This technology combines the best features of both spectrometry and imaging by acquiring both spectral and spatial information of objects. Advances of this technology was in the late 80's/early 90 even though it started in 1972 following the launch of NASAS Earth resources technology satellite [3]. In 1985 became more common to use it in satellites, where it was used for mineral mapping [4] and in the 80's it was deemed feasible to use in crop monitoring [5]. In the late 90's this technology expanded in the agricultural/food industries with the development of the "precision crop management system (PCM)" in 1995 [5] to optimise crop production, some years later in 1998 the technology was used on kiwi fruits to predict soluble solids [6] as an example on application. The technology has grown in popularity ever since especially because of the development in optics and computers. It is expected to replace conventional spectrometry and imaging technologies [2].

Now, with every other emerging technology, it is important to test its ability to accurately measure properties of one its main purposes, food products. It has been well known that properties of food vary widely, internally and externally ex. Wheat kernels form the same sort have variation regarding moisture and protein content. It is therefore an evident possibility to test this state-on-the-art technology on wheat, which varies widely in properties, both internally and from kernel to kernel. So, in the next chapter there will be an overview of wheats significance in human history and what properties are important in order to grade wheat.

1.1 A Brief History of Wheat

Wheat is one of the most influential cereal crops that has existed for consumption for more than 10 000 years with 600 million tonnes being harvested in late 2000's. It is considered to be in the "big three" of cereal crops together with rice and maize. In 2007, 607 million tons of wheat, 652 tons of rice and 785 million tons of maize were harvested across the world. However, wheat is unmatched regarding its cultivation and diversity. It can be cultivated in Scandinavia and Russia at 67° north to Argentina at 45° south to some elevated regions in the tropics. Moreover, wheat has acquired great significance in regard to culture and religion. The wheat is ex. used in the holiday of the "Jewish Passover" in the form of the traditional matzo bread and is also used in the bread to represent the body of Jesus Christ at the Holy Communion, one of the well-known sacraments of Christianity and is seen as a sacred food in some communities of Central Asia and Middle East [7].

There's an indication that wheat originated from the south-eastern part of Tukey and was first cultivated 10000 years ago in the first agricultural revolution ("The Neolithic Revolution") and that cultivation spread it to a major part of the Middle East around 9000 years ago. In the ancient times, farmers choose the wheat that gave superior yield from wild populations. This domestication gave specific genetic traits such as improved wheat spike durability and "easier-to-access" grains as the so-called free-threshing naked forms came forth, *cf. Figure 1*.

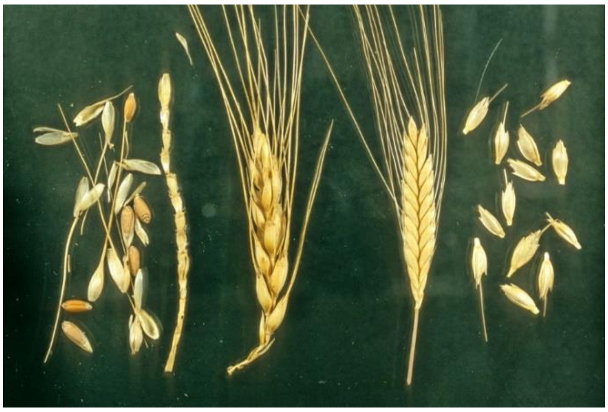


Figure 1: Left, naked wheat. Right, hulled wheat. By Mark Nesbitt [8].

Furthermore, wheat has a very high yield, as one hectare can yield up to 10 tons in a perfect environment. Realistically however, this number is set to a global average of 2.8 tons per hectare since the environment isn't perfect with regard to nutrients and pathogens. However, high adaptability of wheat and high yields are not the only criteria for the success of wheat throughout history. A key factor is that wheat flours can produce doughs with unique properties that can be used to produce a large range of baked goods (breads, cakes, biscuits, pasta, noodles). These properties are primary caused by the proteins in the wheat and their interaction [7].

Studies using proteomic analysis have found that a mature wheat grain can contain 1125 individual components, but only a small number of proteins seem to have impact on the qualities of the wheat. The proteins that have the biggest impact on the wheat are the prolamin storage proteins, the gluten proteins (which makes up for about 80% of total protein) [7].

1.2 Review of grading Parameters of Wheat

There are several grading parameters used to grade either grain or the flour. One of the most important for grain grading is the kernel hardness which directly influences end-use and the milling process. It is mainly measured as a mechanical property of the endosperm more specifically as "the resistance to crushing". The methods to measure hardness will be described in later chapters. Because of hardness, one can classify wheat into three classes, soft wheat, hard wheat and very hard durum wheat [9].

Protein content in wheat is another key indicator of quality and even determines the market value making the quantitative analysis of protein of upmost importance. The quantitative analysis is even a necessity within food labelling and quality control [7], [10]. The protein content can normally be related to the type of wheat. The soft wheat normally has a protein content between 8–11 % while hard wheat and durum wheat have protein ranges from 10–14 % and 9–18 % respectively [9].

The end-use of wheat are dependent on the parameters, hardness and protein content. The end-use can roughly be summarized in the following figure, *cf. Figure 2*.

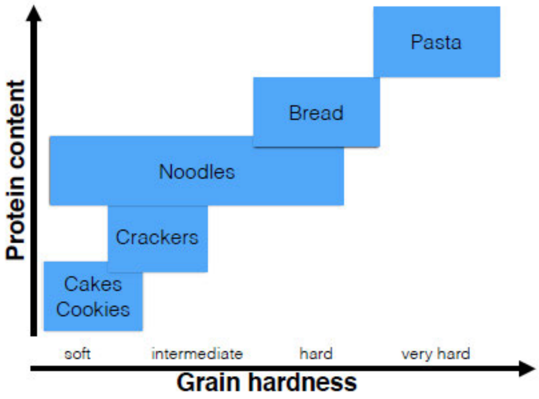


Figure 2: End-use of wheat dependent of protein content and hardness. Adapted from [9].

Another important property of wheat is the optical parameter of vitreousness/vitreosity which is determined visually. The parameter resembles if a kernel has a translucent or glassy appearance.

The connection between hardness, protein content and vitreousness is that vitreous kernels tend to have high protein content and high hardness while non-vitreous kernels are soft and have low protein content. The vitreous property occurs because of lack of air within the kernels making the kernel less dense. Difference in air content happens doing drying of the wheat where the protein shrinks, followed by a rupture of the protein, leaving air spaces inside the kernel making it non-vitreous. In a vitreous kernel there is a lack of rupture doing shrinking [11].

Other parameters for wheat are but not limited to the following, *Table 1*.

Table 1: Review of other quality parameters of wheat. [9]

Grading parameter	Description
Test weight	Indicates the quality of wheat and this parameter is sensitive to moisture content, the shape of grains and the number of broken kernels. It is normally the weight of 100 kg of wheat kernels. It can also be expressed as thousand kernel weight.
Break flour yield	Resembles how much of the flour is recovered during milling of the wheat kernels.
Starch damage	Resembles how much of the starch is damaged. When starch is damaged it has a tendency to absorb more water influencing the dough.
Colour	Wheat can normally be classified as red wheat or white wheat. The white wheat is normally used for Asian noodles [12].
Moisture content	This is important to take into account since high moisture content promotes the formation of micro-organisms in the wheat [13].

Studies have been conducted with hyperspectral imaging on the three quality factors protein content, hardness and vitreousness [14]–[17]. It would therefore be of great interest if it was possible to measure all three quality parameters on the same kernels and hereby make a model or models that can predict all parameters simultaneously with good prediction. This will be the main goal of this master thesis.

2 Problem Statement

The main goal of this master thesis was to examine the feasibility of using hyper spectral images in the near infrared range as a rapid and non-destructive method for determination of multiple quality parameters such as protein content, wheat hardness and vitreousness for one big collection of kernels. This was done by generating multiple models that would allow prediction of said quality parameters with only spectroscopic data. The further usage of these models would be to implant them for quick separation of wheat kernels and making a foundation for a repaid in-line prediction of wheat kernels on a singular basis if the accuracy would allow it.

The quality parameters chosen for the wheat were protein content, hardness and vitreousness. These quality parameters were selected for three main reasons; Mainly, because of their importance for wheat quality assessment, because of the possibility to analyse them in a semi non-destructive manner and because other studies have been conducted with these parameters with well proven reference methods. One can hereby compare the results of this master thesis to the results of the studies. Furthermore, the classic methods to predict protein content in wheat tend to be time consuming and tedious. A great advantage in the industry would be to develop such a state-of-the-art in-line method for rapid protein prediction of the wheat kernels to sort wheat regarding one of wheats most useful quality parameters, the protein content.

To be able to achieve the aim of the thesis following objectives have been defined:

1. Consider reliable, precise and time-saving reference methods for determination of the quality parameters, protein content, hardness and vitreousness based on local equipment at Aalborg university Esbjerg. Criteria should be that only one of the methods are destructive so all parameters can be measured on the same kernels.

2. Testing and validating the used methods comparing the results with known properties of the wheat kernels e.g. average content of a given quality parameter.

3. Acquire of hyperspectral images for the kernels at different orientations and carry out the image analysis. Afterwards, calibrate prediction models for the three quality parameters and evaluate what model is the best under different chosen conditions in pre-processing, variable selection or orientation. A comparison will be done with available literature.

4. Judge, if any other mathematical method were useful for producing prediction models and consider a possibility to use the same method for detection of sick or damaged kernels (which should appear as outliers in the generated models). This judgement will be based on other studies measuring the same quality parameters.

3 Theoretical Foundation for Reference Methods

In this chapter, there will first be a description of the most common methods used for determination of the wheat qualities protein content, hardness and vitreousness of wheat. Afterwards a thorough review on the principles of near-infrared spectroscopy and hyperspectral images will be presented.

3.1 Protein Determination

There exist numerous methods for determining protein content in wheat kernels (and other foods). Some of the most well-known are the Kjeldahl protein analysis and the Dumas protein analysis. These methods will be described in more detail while the lesser known only will be mentioned here, some of them are the use of Nessler regent, Biuret method, Dye binding, Lowry method, Bradford method. Many of these methods require that a given objects protein need to be solubilized before measuring the protein content.

3.1.1 Kjeldahl Protein Analysis

The Kjeldahl method introduced by Johan Gustav Cristoffer Thorsager Kjeldahl, first made public in 1883, is still in use in modern times for the indirect measurement of protein. Specifically, it measures nitrogen. It was first developed for the brewing industry to follow proteins development in the fermentation and germination processes in grain. It was advantageous to measure protein in this industry since a lower amount of protein could mean a higher volume of beer [10].

The structure of protein would not have influence on the estimation since the method only “looks” for total organic nitrogen content. As a limit, Kjeldahl’s method does not distinguish nitrogen in protein and nitrogen from other sources. This can lead to overestimation of the protein content. The other species of nitrogen could come from ammonia or urea [10].

When estimating the protein content in foods based on nitrogen determining methods it is necessary to multiply the nitrogen content with a so-called nitrogen-to-protein conversion factor. Historically, this value has been set to 6.25, based on an assumption that proteins in average contains 16% nitrogen. However, other studies suggested that because of the other nitrogen containing compounds in foods this ratio is different from foodstuff to foodstuff. 6.25 is therefore only a “crude estimate” and when analysing foods, one must choose their unique ratio. Ex. for whole wheat, this value is 5.83 and for rice and corn, 5.95 and 6.25 respectively [18]. Moreover, since the amino acid composition ranges from 13.4–19.3 % in foods, the general 6.25 ratio (16% nitrogen) only seem more inaccurate [10].

The procedure of Kjeldahl can be summarized in three steps:

1. In the acidic digestion step, one uses concentrated sulfuric acid to digest a nitrogen-containing sample in an aqueous environment. The nitrogen will be converted into NH_4^+ ions by the following reaction train, *cf. equation* 1[10][19].

   $$\mathrm{NH_2(CH_2)_pCOOH + (q+1)H_2SO_4 \rightarrow (p+1)CO_2 + qSO_2 + pH_2O + [NH_4^+][HSO_4^-]} \qquad 1$$

2. The next step, distillation, involves transformation of NH_4^+ into NH_3 by addition of an alkali (ex. $NaOH$). With NaOH the reaction goes as the following, *cf. equation 2*. [19]

   $$\mathrm{[NH_4^+][HSO_4^-] + 2\,[Na^+][OH^-] \rightarrow NH_3(g) + 2H_2O\ (a) + Na_2SO_4} \qquad 2$$

   The NH_3 is then collected in the distillation end flask.

3. In the final step boric acid (H_3BO_3) is used for direct titration, *cf. equation* 3 *and* 4 [19]. Firstly, the ammonia is captured in excess H_3BO_3 and makes an ammonium-borate compound. A colour change will happen followed by the rising of the pH. Afterwards this compound will be directly titrated by use of a mineral acid, can be chloric acid HCl. One mole of HCl corresponds to one mole of NH_3. The H_3BO_3 will be to weak an acid to disrupt the titration with HCl [20]. "A reverse" colour change will happen when the compound has been neutralized.

   $$\mathrm{NH_3 + H_3BO_3 \rightarrow NH_4H_2BO_3 \rightarrow NH_4^+ : H_2BO_3^-} \qquad 3$$

   $$\mathrm{NH_4^+ : H_2BO_3^- + HCl \rightarrow H_3BO_3 + NH_4Cl} \qquad 4$$

   Knowing the amount of chloric acid (HCl) used to neutralize the ammonium-borate, with combination of a proper indicator, one can calculate the nitrogen content and afterwards by using the nitrogen-to-protein ratio calculate protein.

One of the main disadvantages of the method is that it is quite time consuming. The titration and distillation can be completed in 15 minutes, the digestion varies greatly depending on the sample with time reaching from 30 minutes to several hours. Furthermore, in the original method, samples where as large as 1–2 g [10] making measurement of single kernel wheat impractical. However micro Kjeldahl methods exist, but the proposed time for completion is still 2 hours and this method also requires specialized sand beds reaching temperatures of 380˚ C [21]. Other disadvantages include that the method needs handling of sulfuric acid, heavy metal catalysts and the method is overall labour intensive. The advantages of Kjeldahl is its precision in measurement of protein, non-dependent of the samples physical state. Moreover, the Kjeldahl method only requires cheap equipment [10].

Regarding the digestion procedure of Kjeldahl an alternative digestion exists. This alternative method includes the oxidation of all organic and inorganic forms of nitrogen using peroxydisulfate $S_2O_8^{-2}$, in the form of potassium persulfate $K_2S_2O_8$ in an alkaline environment (the source for this paragraph used NaOH [22]) at high temperatures and pressures. This digestion, done in an autoclave, should oxidize all nitrogen-containing compounds to nitrate (NO_3^-) in a water sample [23]. So, in contrast to the original method, it is NO_3^- that is the focus of digestion instead of NH_4^+. The concentration of nitrate is then seen as the equivalent to the total nitrogen amount. The study [22] could confirm very good nitrogen recovery when analysing both organic and inorganic nitrogen for aqueous standards. However, they experienced that the standard protocol of autoclavation was insufficient and their samples were not fully digested at 121 °C in 30 minutes. They suggested to modify the time of autoclavation up to 60 minutes [22]. After digestion other methods are required that determines nitrogen based on nitrate. This digestion procedure are normally done on water samples with suspended solids. [23], [24] . A method to complete this task and analyse the nitrate can ex. be done with colorimetric flow injection analysis.

3.1.2 Dumas Protein Analysis

Another method for analysis of protein is the dumas combustion method. Compared to Kjeldahl this method is less tedious and time-consuming because of the opportunity to use automated and easy-to-use laboratory equipment. This enables that a sample can be analysed in approximately 6 minutes. The foundation of this method is to use an induction furnace to make a complete combustion of all forms of nitrogen to generate nitrogen oxides (NO_x) at temperatures ranging up to 1050–1300 °C. Other combustion gases include O_2, CO_2, H_2O, N_2. The unwanted gases are removed with traps so only nitrogen oxides and nitrogen are left. Afterwards these nitrogen oxides are reduced to N_2. Finally by measuring the thermal conductivity, one can quantify the amount of N_2 [25][26]. This method therefore removes the necessity to preform wet chemistry analysis on the samples and this method may have improved repeatability and accuracy compared to Kjeldahl [25]. A study [25] compared the ratios of crude protein between the method of Kjeldahl and Dumas for different food products, *cf. Table 2.*

Table 2: Ratios between Kjeldahl and Dumas method for different food products. [25]

Food product	Ratio (Kjeldahl crude protein / Dumas crude protein)
Dairy	1.01
Oilseeds	1.00
Infant formulas	0.98
Cereals	0.95
Meats	0.94
Vegetables	0.89
Fish	0.80
Fruits	0.73

Dumas gives in general a higher estimation of protein (around 1.6% [27]) because it determines total nitrogen in comparison to Kjeldahl where the assumption is normally that inorganic species are negligible. However, when using proper constants the study suggested that the Dumas method could replace the Kjeldahl method for protein analysis in selected food groups [25].

To sum up on the advantages of Dumas compared to Kjeldahl, Dumas is deemed more environment friendly (less use of chemicals), safer, easier method to implant and a faster method. However, the main disadvantage of Dumas is that the equipment used is expensive to buy and requires lots of experience to maintain properly making Kjeldahl the more economical but time consuming choice [28].

3.1.3 Flow Injection Analysis

Another possibility to analyse protein content, if Dumas equipment is not available and the Kjeldahl method is considered too time consuming, is the use of flow injection analysis (FIA). There exits to sub-methods of FIA: gas diffusion FIA or colorimetric FIA. FIA (based on gas diffusion) keeps the digestion of the Kjeldahl method but automates the distillation and the titration (determination) step which greatly reduces analysis time and chemicals used [28]. The main goal of ("colorimetric") FIA is to automatically through a series of reactions make a samples chemical of interest into a colour compound that can be analysed though spectroscopic means. The mechanism and basic components of a FIA apparatus starts with a stored reagent or multiple reactants that also acts as the carrier solution. These reactants are propelled with constant flow through a series of tubing's and is mixed with one's sample. After reaction one would generate the colour compound [29]. Specifically, for indirect measurement of protein by nitrogen determination the main principle are as follows as stated in AN 5202, *Determination of total oxidized nitrogen in water by FIASTAR 5000*:

- After digestion of the sample, the sample passes through a cadmium reductor to reduce the nitrate to nitrite. So here it is necessary to use the alternative digestion procedure to generate nitrate.

- The formed nitrite will with addition of an acidic sulphanilamide solution form a diazo compound.

- This compound is reacted with N-(1-napthyl)-Ethylene Diamine Dihydrochloride (NED-reagent) to generate a purple azo dye.

- This purple dye can be measured at a wavelength of 540 nm.

A study [30] used gas diffusion flow injection analysis on dairy products (cheese, milk, yogurt), they could achieve sampling of 100 per hour with good precision excluding digestion time. The relative difference between their Kjeldahl reference values and obtained values using FIA were less than 4 %. They found no significant difference between the two methods at a 95% confidence interval [30]. Other studies exist that also have used gas diffusion FIA [31]–[33] with great precision and accuracy. A study [34] using the colorimetric version of FIA suggested that the use of FIA could be used as a suitable method for measurement an indirect measurement of protein. On a side note, when measuring the final product, a coloured compound from the protein digestion reaction train, it is possible that the colour reaction did not reach completion. Temperature and flowrate are therefore of crucial importance for reproducibility did the study note [34].

Their results could be obtained within a minute with sample sizes ranging from 5–20 µg with a linear response in the protein content region from 0–10 µg. For them, it was possible to run 60 samples an hour.

FIA is also deemed competitive from another study [28] regarding simple setup, precision and sensitivity. Based on the ranges and the positive onlook, FIA could be a suitable method for the analysis of single wheat kernels that is worth looking into. Therefore for timesaving and convenience reasons the colorimetric version of FIA will be analysed in preliminary experiments to tests its ability to predict protein.

3.2 Hardness Determination

One can define hardness of wheat as *"resistance to deformation"* and it is determined under four stress principles namely, tension, compression, shearing and bending ([11], [35]). Factors that influence the hardness include protein, moisture content, interaction between the two together with interaction between starch and minerals in the endosperm matrix. The first methods to measure wheat kernel hardness mechanically was developed in the early nineteenth century. It measured the force required to crush a kernel [35]. Nowadays, there exists over 100 methods for measuring wheat kernel hardness [36], therefore only the most common ones will be mentioned. Common methods of measuring wheat hardness, approved by the American association for clinical chemistry (AACC), are *the single kernel characterization system* (SKCS), particle size index (PSI) and Near-infrared spectrometry (NIR). The principles of the SKCS apparatus are that wheat kernels are crushed between two mechanical moving parts with a calibrated force cell attached. This method also measures weight, diameter and moisture all of which must be calibrated with a reference (laboratory) method. Weight is measured by measuring the electrical force required to bring a placeholder to the origin position. Diameter is measured as the gap between the two crushing parts at the point of contact. Moisture content is based on electrical conductance change under maximum crushing force. The hardness index for this method can be calibrated by having two sets of wheat classified as soft and hard from the wheat provider. Soft get their hardness index defined as 0 and hard get their index defined as 100. Afterwards a linear discriminant analysis can be made to make a linear discriminant equation, the results coming from new kernels can then be expressed in values from 0–100 and depending on the placement on the scale, they can be classified as soft, mixed or hard wheat sorts [37].

The particle size index uses a representative sample of wheat that is initially weighed. Afterwards, the sample is grounded, and the resulting flour is transferred to a sieve and shaken for a specific amount of time. The index is then calculated by expressing the passed-through flour weight (w_1) as a percentage of the original flour weight (w_2) [36].

$$PSI\% = \frac{w_1}{w_2} \cdot 100\%$$

Near-infrared spectroscopy (NIRS) has been a promising technology to use when predicting hardness of wheat, a study [38], used NIRS with its reference method in SKCS to accurately predict hardness of different wheat species ($R^2 = 0.91$ for a 30 sample, SECV = 7.70) in the band region of 950–1690 nm. NIR can therefore be a rapid non-destructive measurement of wheat hardness [38].

Many methods for measuring hardness of wheat are destructive in nature, however, a semi-destructive method exists that measures the rupture force of wheat kernels as an indication of hardness [39]. Here a simple compression device is used to compress the kernels. In a recent study [39], it was found that the correlation levels between protein content and hardness measured in this way was as high as r = 0.953. Making this an attractive method to evaluate.

3.3 Vitreousness Determination

The main methods to measure vitreousness in literature are by the international standard 129 and a modified version that splits the wheat into more vitreousness classes [40]. In the international standard method, a sample consisting of 100 g of kernels are divided into two classes. Either they are defined as fully vitreous or others-class (which includes kernels not-fully-vitreous or damaged kernels). To be defined as fully vitreous a kernel should have complete absence of mealy appearance. After division of the sample, each class group gets weighed. The vitreousness is then determined to be the percentage of fully-vitreous kernels. The alternative method includes a special cutting-device called a farinator, *cf. Figure 3*. This device can hold up to 50 kernels, firmly, and cut them in half transversely with a blade. This would now give 50 x 2 samples which visually can be evaluated. This evaluation is based on classification into four groups [41].

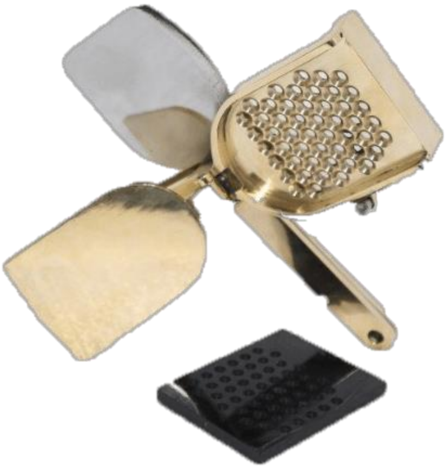


Figure 3: Showcase of a farinator. [42]

- Group A) this group consists of kernels that are fully vitreous.
- Group B) this group consists of kernels which cross-section is 75–100% vitreous.
- Group C) this group consists of kernels which cross-section is 50–75% vitreous.
- Group D) this group consists of kernels which cross-section is 25–50% vitreous.

Now, the vitreousness in percentage is calculated by using the weighted fraction of these groups:

$$\mathrm{V}(\%) = \mathrm{A} + \frac{3}{4} \cdot \mathrm{B} + \frac{1}{2} \cdot \mathrm{C} + \frac{1}{4} \cdot \mathrm{D}$$

Here A, B, C and D are the fractional amount of each kernel in that specific group.

A visual representation of an example of vitreous classing can be reviewed as follows, *cf. Figure 4*.

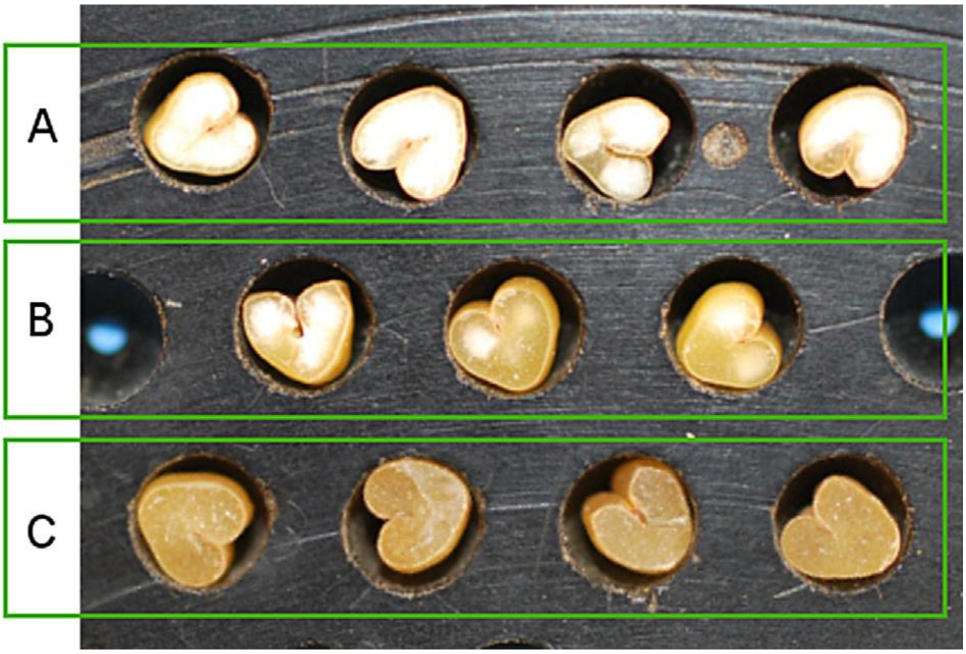


Figure 4: Different classes of vitreous kernels. A is non-vitreous, B is piebald and C is fully vitreous. Permission to use this image has been granted by the source [40].

A study, [40], tested the correlation between these two methods and found a high correlation coefficient of 0.87 indicating that there are not big difference in the vitreousness (%) values from the two methods and both methods are suitable to determine vitreousness.

3.4 Spectroscopic Methods

Here an introduction to the properties of light and their interaction with molecules will be presented to help understand the spectroscopic methods.

Spectrometric methods have been widely known to serve as an accurate method for both qualitative and quantitative analysis of composition and properties of a given object by using the information of light absorption in the electromagnetic spectrum to give a "signature". Near-infrared spectrophotometry (NIR), with reference in another method, allows the analysis of fully intact samples [10]. This makes the use of NIR a very favourable non-destructive method to use when analysing objects in the food/agricultural industries. These objects could well be wheat. The foundation of spectroscopic methods lays in some specific properties of molecules. Molecules have two sets of property classes, static (atomic composition, morphology and stereochemistry to name a few) and dynamic (rotation, vibration and molecular transitions), which is influenced on the static properties. The dynamic properties are also seen as an expression of the energies of the molecule and it is those properties that provides a baseline to make a chemical spectrum of a molecule [43]. The most basic definition of a chemical spectrum is: A view of electromagnetic radiation coming from a molecule. The radiation is a result of the molecules interaction with a light beam. Either it absorbs light in specific frequencies and the transmitted light is detected or, reasoned a shift from a high to a low energy state, a new (emission) radiation is generated.

This phenomenon is caused because of the energy properties of light, light can both be considered as a particle and a wave. The energy of a light particle (a photon) can be described with the following equation and relations [43].

$$E_t = h \cdot v = h \cdot c \cdot w_n = \frac{h \cdot c}{\lambda}$$

$$c = v \cdot \lambda$$

$$w_n = \frac{1}{\lambda}$$

Here E_t is the energy of light, v is frequency of light, h is Planck's constant, c is the speed of light, w_n is the wavenumber, λ is the wavelength of the light. On a side-note, wavenumbers are another representation of wavelength, typically one displays wavelength as wavenumbers based on the following conversion [43]:

$$\text{wavenumbers [cm}^{-1}\text{]} = \frac{10^7}{\text{wavelength [nm]}}$$

3.4.1 Near-infrared Spectroscopy

The foundation of near-infrared spectroscopy (NIRS) is the determination of absorbance of light passed through a sample in the near-infrared region. The near-infrared region is normally defined with wavelengths between 780–2500 nm. This corresponds to wavenumbers in the interval of 12820–4000 cm^{-1} respectively. The absorbance is caused by specific vibration of chemical bonds in a chosen sample. In NIRS the main molecule energy classes of interest is the vibrational and electronic energy forms[1], with vibrational energy being the most predominant [43].

Bonds between a molecule's atoms can oscillate in a specific manner giving rise to the vibrational energy. The energy of such an oscillating system can be defined as [43]:

$$E = h \cdot v_{fund.} = h \cdot \frac{1}{2 \cdot \pi} \cdot \sqrt{\frac{k}{\mu}}$$

$$\mu = \frac{m_1 m_2}{m_1 + m_2}$$

Where h is planks constant, v$_{\mathrm{fund}}$ is the fundamental frequency also shown in the equation, k is the force constant, m$_1$ is the mass of atom 1, m$_2$ is the mass of atom 2. However, to determine the additional discrete quantum energy levels of a unique vibration one includes the vibrational quantum number *v*, which can take the form of whole numbers (v = 0, 1, 2, …). The energy equation adjusts to:

$$E = \left(v + \frac{1}{2}\right) \cdot h \cdot v_{fund.} = \left(v + \frac{1}{2}\right) \cdot h \cdot \frac{1}{2 \cdot \pi} \cdot \sqrt{\frac{k}{\mu}}$$

To review these levels, one can plot the potential energy of the system defined as: $V = \frac{1}{2} \cdot k \cdot x^2$ from equilibrium position (quantum number = 0) against the displacement of the atoms on the x-axis[2], *cf. Figure 5.* For this ideal behaviour there are two assumptions for the quantum numbers *(cf. Figure 5 for reference of quantum numbers.).* 1) There cannot happen a vibrational transition, where the quantum number is greater than one. 2) Some transitions, where the quantum number is one, can also be forbidden based on group theory [43]. A discussion of this mathematical discipline mentioned is not included in this master thesis to avoid getting side-tracked.

[1] If the sample is in gas-phase, then rotational energy also plays a part. [43]

[2] k is here restoring force constant.

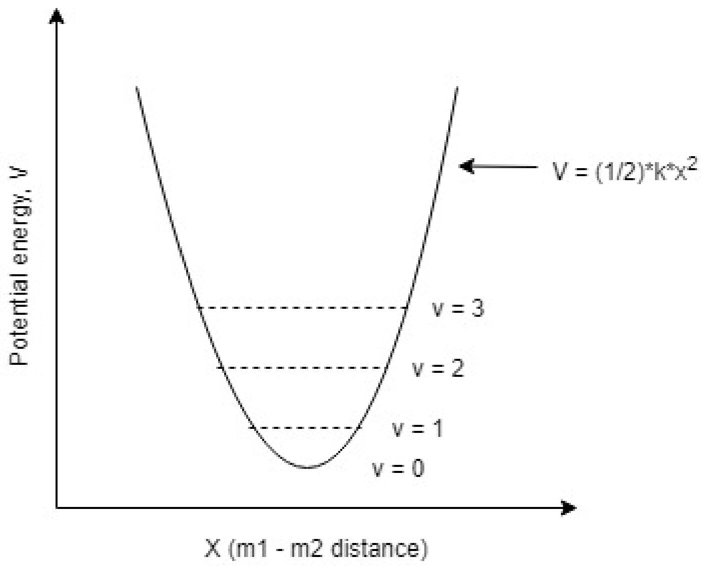


Figure 5: Harmonic oscillation of two atoms. Potential energy against displacement of the two atoms. Figure is self-made.

However, these vibrations are based only on interaction between two atoms in a molecule. For most cases, the vibrations are an interaction between multiple atoms and there exists many different types of vibrations (stretching, bending, rocking, twisting to name a few). Therefore, a new term arises, coupled vibrations, vibrations that are an interplay with other vibrations. A common example of a coupled vibration could come from proteins. Proteins peptide bond can make simultaneous vibrations of the C=O, C-N and N-H bonds. [43]

For a molecule being able to absorb a specific light energy, the lights energy must be equal to the difference in two of the energy levels, therefore one can combine the equation of light energy and the equation of energy quantum levels in an oscillating system to the following condition [43].

$$E_t = \Delta E = E_{v2} - E_{v1}$$

Only specific frequencies or wavelengths of light can be absorbed because the only adjustable parameter[3] in the equation for energy of light is the wavelength/frequency. But one additional condition needs to be satisfied, there is a need for an oscillating electric field, a dipole change of the molecule caused by interaction with light to make the absorption happen. To visualise, a CO_2 molecule can have anti-symmetric stretching of its bonds with oxygen, here a difference in charge density is happening and switching of the concentration of electrons in one oxygen end to the other end causes dipole movement. Absorption will happen. However, the systematic stretching of CO_2 cannot absorb light. Dipole movement has great influence with the intensity of the peaks seen on a chemical spectrum. It can be stated that the magnitude of change of a dipole movement is proportional to a given vibrations intensity in a chemical spectrum [43].

Big, complex, asymmetric molecules experiences small dipole changes in them and got high possibility to absorb light. It is this ability to absorb light at unique wavelengths/frequencies between light and molecules enables the creation and usage of analytical spectroscopy. [43]

[3] On a side note, the speed of light c can also slightly change based on the medium it moves though.

Now it is worth mentioning that almost all vibrational states start in quantum number zero, so they can only move up to quantum state 1 in the ideal cases. However, most functional groups that makes this transition need lights with wavelengths between 2500–25000 nm to “fuel” this process[4]. This is light in the mid-infrared region. To get signals in the NIR, one must take nonideality into account [43].

The theory covered yet has the foundation in ideal behaviour of molecules, but vibrational energy of molecules needs to have more assumptions on nonideal behaviour to reflect real life scenarios. Inclusion of additional assumptions also explains two other important phenomena’s in NIR spectroscopy called combination and overtone bands. The two nonidealities assumptions to include are mechanical and electrical anharmonicity. The assumption of the former anharmonicity, the mechanical anharmonicity, is that most real molecules experience anharmonic vibrations instead of harmonic vibrations and the assumption of the latter, the electrical anharmonicity, is that the magnitude of dipole change and inter atomic distance cannot be a linear function as seen in the ideal world [43]. This will be explained.

Inter atomic distance against the dipole movement is not linear because if the distance between the atoms goes to zero, no dipole movement will happen and if the distance goes to infinity (separation of the atoms) no dipole movement will happen. However, dipole movement happens, so this means that there must be a maximum of dipole movement. This makes the electrical anharmonicity valid. [43]

An example of a harmonic vibration was seen in *Figure 5,* however the potential energy is better approximated using: $V = k_1 \cdot x^2 + k_2 \cdot x^3 + k_3 \cdot x^4 + \cdots$ instead. This is because, when atoms are squeezed together, they experience a strong repelling force and when they move away from each other, they will dissociate at some point, so the oscillation cannot be harmonic. This also makes the mechanical anharmonicity valid. [43]

This nonideal behaviour makes it possible to have: [43]

- Vibrational transition jumps with quantum numbers higher than one.
  So-called *overtone* transitions, seen in the NIR-region.
- Simultaneous increase in multiple quantum numbers (different vibration types) caused by a single photon. So-called *combination* transitions, seen in the NIR-region.
- Different energy-differences for each transition. In ideal cases the energy difference between the quantum states are equal.

Energy levels can now be calculated by the following formulae [43].

$$E = h \cdot v \cdot \left(\boldsymbol{v} + \frac{1}{2}\right) - x_m \cdot h \cdot v \cdot \left(\boldsymbol{v} + \frac{1}{2}\right)^2$$

$$\left(v = \frac{1}{2 \cdot \pi} \cdot \sqrt{\frac{k}{\mu}}\right)$$

[4] This corresponds to wavenumbers of 4000 - 400 cm^{-1}.

The new term introduced here, x_m, is the anharmonicity constant of vibration. It varies depending on type of vibration and type of molecule with a range of 0.005 - 0.05. [43]

The following figure showcases these energy levels and the shape of the potential energy of a diatomic molecule for a nonideal anharmonic oscillator model, *cf. Figure 6.*

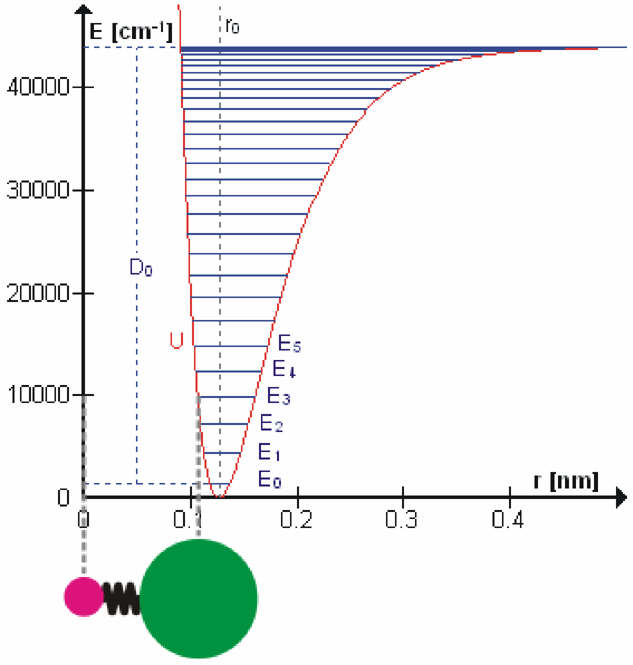


Figure 6: Anharmonic oscillator model. All credits goes to Darekk2 for this picture under the creative commons license. [44]

Figure 6 showcases a HCl molecule making an anharmonic vibration in energy level three (E_3), where energy is given in wavenumbers, U is here potential energy, D_o is dissociation energy, r_0 is the equilibrium bond length. One can see that the molecule is in a compressed state. The formulae for the potential energy of this system (anharmonic vibration) is given by the following equation with definitions as used in the picture. [43]

$$U = D_0 \cdot \left(1 - \exp\left(-a \cdot (r - r_0)\right)\right)^2$$

The new term here is a, which is a constant, which magnitude relates with specific electronic states of given molecules. Still, this showcasing corresponds to a diatomic pair, for polyatomic molecules the vibrational energy levels gets even more complicated and includes additional interaction terms, to review it please see chapter *Appendix of theoretical interest* However, it is this interaction that enables the overtones and combination bands seen in the NIR region [43]. Classic overtones bands seen in the NIR region comes from functional groups such as CH, NH and OH.

However, it is stated that three other conditions for combination bands needs to be satisfied [43]:

- The vibrations that are to be combined must have the same functional subgroup.
  A good example is the -CH_2-
- The vibrations of a combination must be symmetric.
  A good example is when the -CH_2- makes symmetric stretch and bending in the same plane. Symmetric stretch can be found at ~3400 nm (2870 cm^{-1}), the bending can be found at ~6840 nm (1460 cm^{-1}). The combination of these two can be found at ~2309 nm (4330 cm^{-1}, in the NIR).
- No combination of other overtone signals.

The intensity of these combination and overtone peaks in the NIR depends on the degree of anharmonic vibration and dipole movement. Higher degree of anharmonic vibration or dipole movement gives higher signals. [43]

The most useful absorption bands in the region of NIR comes from molecular vibrations from C-H, N-H, O-H and S-H groups making the method of NIR efficient for analysing biochemical species. The following image showcases the placement of overtones and combinations bands of these most common groups. All rights reserved to Metrohm. [45]

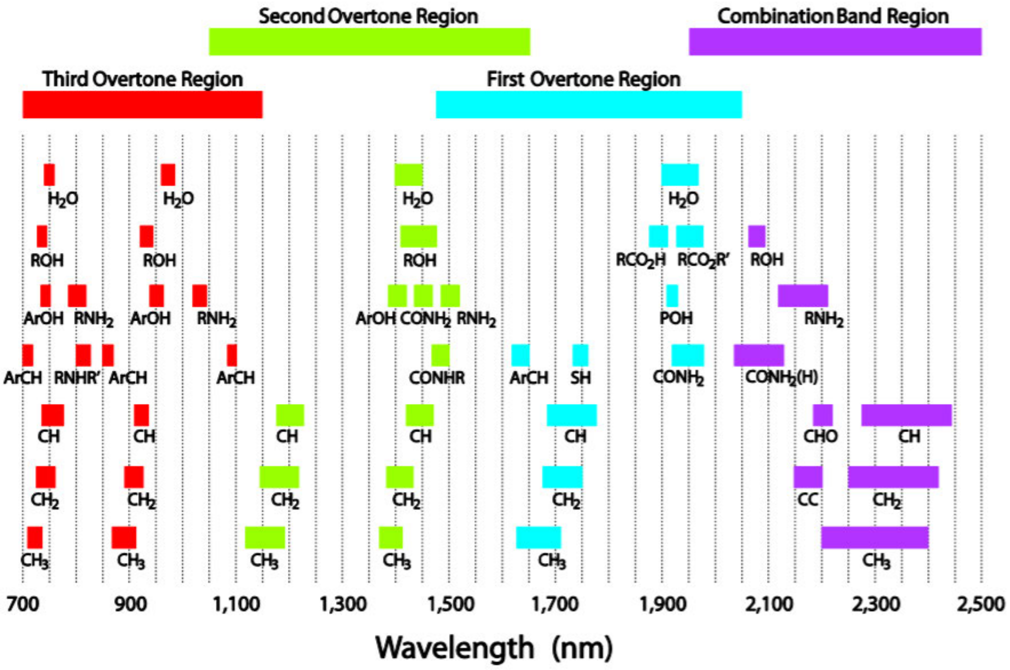


Figure 7: Most important absorption bands for NIR. [45]

Lastly, with calibration in a reference method the method of NIR spectrometry is enabling the detection of specific compounds such as protein with functional groups such as CH, NH and OH. The greatest advantage, of NIRS as mentioned before, is its ability to avoid destructive treatment of one's sample by avoiding extraction or use of other chemical reagents. With a good foundation of data in the reference method and its calibration, this method is deemed precise and can analyse new samples in a very short span of time. [28]

3.4.2 Introduction to Imaging

The foundation of lays in images. Images are two-dimensional[5] representations of physical objects. Images can be represented numerical and, in these cases, the image is called a digital image. Digital images hold geometric (spatial) information and normally have a binary nature. Hyperspectral images (HSI) contain both the spatial and spectral information. There are two classes of digital images, vector class or raster class (also referred to as bitmaps). Vector classes are digital images used in computer graphics, they are made based on mathematics (lines, points, curves) including multiple vectors with different properties on a (x, y)-plane. A vector images size can be altered without losing quality. Raster images, on the other hand, are images that consists of discrete elements. These elements are called pixels, which has variating intensity regarding colour on an image. The quality of raster images can be described in terms of resolution, which is a measure of how many pixels can be displayed on each dimension of such an image. Raster images are therefore resolution dependent, vector classes are resolution independent. [2]

To create raster images, one can use the technique of digitization, but firstly, one must know the basic principles of cameras. This is done based on the pinhole camera model to understand the perspective transformation. The following figure, *Figure 8*, shows the geometry of this model.

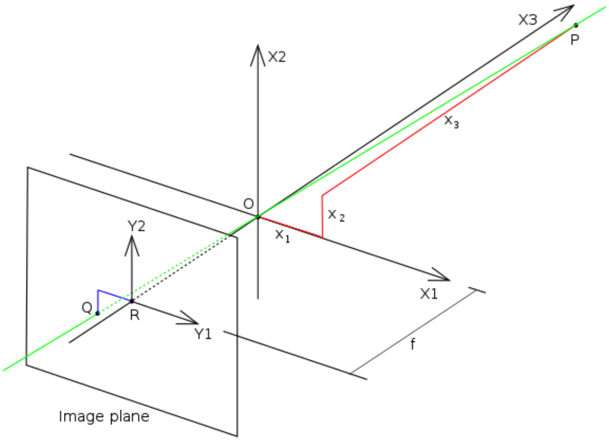


Figure 8: Geometry of pinhole camera. Credits go to Wikipedia user, N3bulous, for issuing this image in the public domain. [46]

The pinhole is in the origin (O) of this three-dimensional system (X1, X2, X3). The pinhole can also be referred to as the lens. The X3 axis is the optical axis, in other words, viewing direction of the camera. "Inside the camera" is the image plane, a 2D plane with its axis being called (Y1, Y2), these are parallel to the axes (X1, X2). The distance from the pinhole (O) to the image centre (R) is called f (negative direction of the X3 axis), normally called focal length. In optics, focal length determines what the field of view is for a given lens. Now, point P, is a random point in the real world with coordinates (x_1, x_2, x_3) in the cameras field of view.

[5] It can also be three-dimensional, such as holograms.

The green line represents the projection of point P into the image plan at point Q with coordinates (y_1, y_2). The line also has intersection in point O. Seen from the X2-axis, “from above”, the geometry looks as following, *cf. Figure 9*.

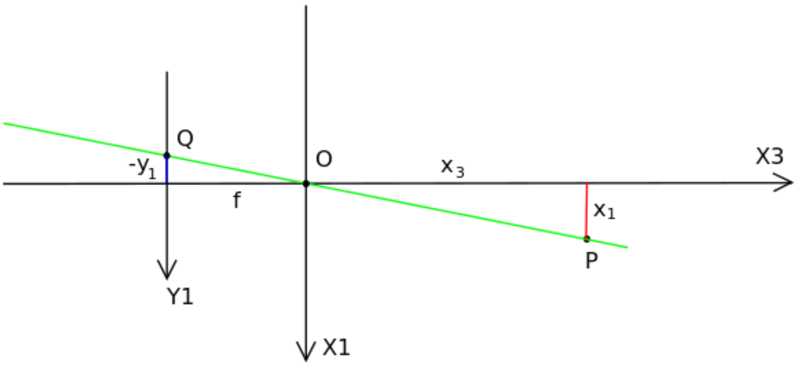


Figure 9: Geometry of the pinhole camera down the X2 axis. Credits to Wikipedia user, KYN, for issuing this image in the public domain. [47]

Now, it is possible to explain the perspective transformation, from 3D to 2D with the help of simple geometry and triangle perspectives. Two similar triangles can be seen in *Figure 9* (sides -y_1 and f for the first and x_3 and x_1 for the second). Since they are similar (because of the intersection in O), the ratios of their sides must be the same. Therefore:

$$\frac{-y_1}{f} = \frac{x_1}{x_3} \leftrightarrow y_1 = -f \cdot \frac{x_1}{x_3}$$

To find y_2 one must look at the triangles down the X_1 axis. The first triangle will have the sides X_2 and X_3 and the second, in the 2D plan, will have the sides f and - y_2 therefore:

$$-\frac{y_2}{f} = \frac{x_2}{x_3} \leftrightarrow y_2 = -f \cdot \frac{x_2}{x_3}$$

In vector notation this resembles to:

$$\binom{y_1}{y_2} = -\frac{f}{x_3} \cdot \binom{x_1}{x_2}$$

Which is the formulae for perspective transformation from 3D to 2D.

Now, the cameras projection is in its foundation, quote, a “two-dimensional, time-dependent, continuous distribution of light energy” [48]. Here the technique of digitization comes into use to convert this type of signal to something useful. Digitization in its core transforms a continuous signal, an analog signal, to a discrete signal, a digital signal. This is done by arranging individual sensor elements on the image’s “sensor” plane. Each of these sensor elements measures the incoming light on them by triggering an electric charge caused by the photons in the time of exposure. This charge build-up can be converted to a finite set of values, which can be represented by a computer as a 2D matrix. These elements, these pixels, can be represented binary, grayscale or in colour. In a colour camera however, a prism is present that splits light into three components red, green and blue light. By using these three matrixes one can build the “real” colour. The digital signals can be represented in an useful form on a monitor ([2], [48]).

To determine image quality, one can access two parameters: spatial resolution and pixel bit depth (colour). Spatial resolution is defined as the frequency of pixels used to describe the scenery space of an image with specific dimensions, it can obtain values such as 640x480 or 2048x1536. The more pixels one has, the higher resolution. Pixel bit depth is defined as the amount of shades a given pixel can undertake based on the information of the digital image. A pixel contains strings of zeros and ones, 1-bit can only assign 0 or 1 to a single pixel, so only two shades can be present. However, an 8-bit system $(2^8 = 256)$ can assign 256 shades. So, higher bit depth gives higher levels of change on an image [2]. One can also improve the quality of an image by manually using pre-processing techniques and these techniques are worth mentioning. But first, an explanation of the principles of the hyperspectral imaging procedure is presented.

3.4.3 Principles of Hyperspectral Imaging

In comparison to the 2D images the hyperspectral images are 3D with two spatial dimensions and one spectral dimension giving hypercubes with the format (x, y, λ). The spectral range can vary depending on application and the system can be adjusted to use either transmission, reflection or emission radiation of a given specimen [49]. One common way for acquiring a hypercube is spatial scanning, *cf. Figure 10*. There also exists other scanning techniques such as snap-shot scanning giving the whole hypercube at once or spectral scanning which scans only one wavelength at a time using different filters with a stationary specimen [49].

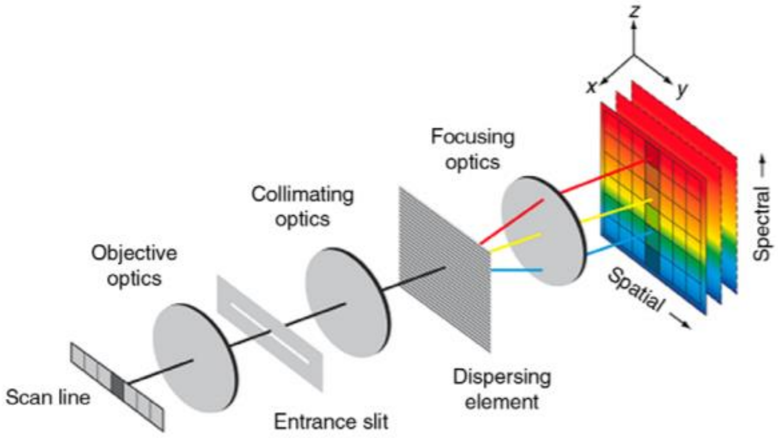


Figure 10: Principles of an image spectrometer (push broom). Permission to use this image granted by the source [50].

In spatial scanning systems initially only a line of pixels in the scene is scanned by limiting the incoming light from the scene to only an entrance slit. The output will therefore only be 2D at first (x, λ), however, by incorporating moving mechanical parts multiple lines can be acquired in the y-axis. There are two available types of scanners available for the line systems, push broom or whisk broom scanners. The more expensive whisk broom scanners have their internal mirrors aligned and moving in such a way that only one pixel is collected at a time in the scene-line typically giving higher resolution of the image. The push broom scanners on the other hand measures all pixels in the line simultaneously. [49]

3.4.4 Image Processing

An important aspect of image quality is the use of image processing. Image processing are multiple methods that are used before image analysis to manually improve the quality of an image by removing imperfections. Imperfections can include, but are not limited to, camera motion, bad focus, noise, nonuniform lightning. The image processing can be classed as follows: ([2],[48])

- Point operations [2]
  The operations are done on a single pixel basis with no regard for adjacent pixels. A function is applied to each pixel, that based on intensity of the given pixel, changes the value of the pixel or keeps the value the same for the enhanced image. This is useful when one wants to separate background and foreground more clearly.

  Common methods include, thresholding, adaptive thresholding (threshold values are chosen locally for separation here), stretching (improves image by "stretching" the intensity range of the pixels).

- Adjacent operations (mask operations) - The usage of filters [48]
  The operations are done on multiple pixel basis; the enhanced pixels depend on the properties of a set of chosen pixels or adjacent pixels. One classic example is image smoothing to remove noise from an image. Here a pixel can be replaced by the average of other adjacent or neighbouring pixels. In these types of filter operations, the size parameter is of importance. Here size is defined as the number of pixels used in the operation. When a filter has a weighted summation of pixels, the filter can be said to be linear. The weight matrix of such a linear filter for a given image is called a filter mask and when this matrix is applied to any given pixel, one can enhance the selected pixel.

  A subclass of adjacent operations is the morphology-operations, which are worth looking more into. A whole chapter will be dedicated to these operations moving forward.

- Global operations [2]
  The operations are done based on all pixels on a given image. Based on the properties of the pixels an "image mask" for the whole image can be defined based on a feature of the pixels. When applied to the image this mask would define pixels in a binary fashion, ex. It sets pixels outside the limits of a feature to zero and pixels inside this limit to one. This makes the uses of masks an efficient way to remove the background from an image.

3.4.4.1 Morphology Operations

Morphology operations or morphology filters are methods that alters a local pixel isle in an image in a predictable way normally done for binary images. The motivation to use morphology operations are to remove smaller structures ex. points or thin lines that clearly are bad noise pixels. The most basic operations are shrinkage and growing of the structures in a binary image. The principle are as follows: All structures on a binary image have one layer or more layers of pixels removed (shrinkage). The smaller structures will disappear while bigger would remain. Now, by growing the remaining structures back to their original size by adding layer of pixels around the border of the structures, the smaller structures are now effectively removed. The two mentioned methods “growing” and “shrinkage” goes under the term “dilation” and “erosion” respectively [48]. These will be described shortly. When explaining these definitions, it is important to include the definition of pixel connectivity. The following image, *cf. Figure 11*, showcases two cases of adjacent pixels.

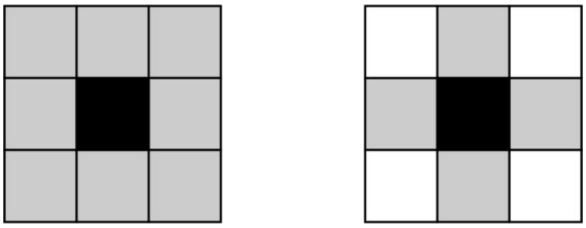
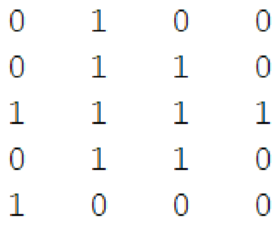
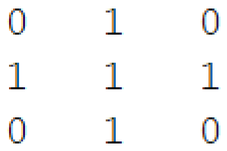
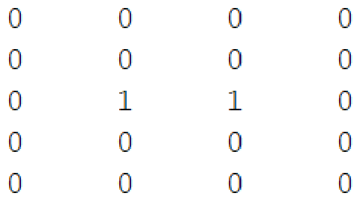
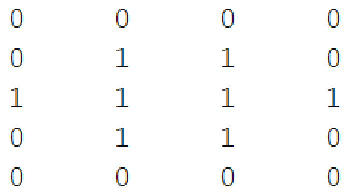


Figure 11: Representation of two types of adjacent pixel types. Thanks to Wikipedia user Master Uegly for placing the image under the public domain [51].

Here two types of pixel connectivity’s are included the 8-pixel “neighbourhood” displayed to the left and the 4-pixel neighbourhood displayed to the right. It is also possible to have a 6-pixel “neighbourhood” on a hexagonal pixel grid. Let’s look at the 4-pixel neighbourhood and the morphology operation of erosion. One can define this 3x3 structure as a structuring element. The structuring element has origin in the black spot (the centre). This structuring element is applied to a binary image by placing the centre on every pixel, if the structure fits on the image (the grey spots align to the 1’s of the binary image), the centre is kept and if not, the pixels are removed in the grey areas under the erosion procedure.

A simple illustration of erosion and dilation made in Matlab illustrates this on the next page, *cf. Table 3*.

Table 3: Simple illustration of erosion and dilation of a binary image.

To the right is an example of a binary image made in Matlab.	 0 1 0 0 0 1 1 0 1 1 1 1 0 1 1 0 1 0 0 0
The structuring element is chosen as the disk with 1-pixel radius. The structuring element can take many other forms and shapes as well.	 0 1 0 1 1 1 0 1 0
When applying erosion with this disk, the centre only fits two places in the original image. The rest of the pixels are removed. To the right is the eroded image.	 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0
Now when applying dilation on this altered image with the same structuring element the centre is moved to every remaining 1-value and the structuring elements adjacent values are placed in the image. The diluted image is as follows. Erosion followed by dilation is also called *opening*. When comparing this image with the original one, one can see that the "noise" pixels are removed while keeping the original structure intact. This example showcases why morphology operations can improve image quality.	 0 0 0 0 0 1 1 0 1 1 1 1 0 1 1 0 0 0 0 0

4 Mathematical Foundation

To help analyse the hyperspectral images, containing several thousands of pixels each with their dataset of wavelengths, the need for chemometrics arises. Chemometrics is in its core a mathematical chemical discipline that uses mathematical and statistical methods for evaluating complex chemical systems by extracting the most important parameters out of the large dataset and finds the most optimal procedure for experimentation.

The first concept in chemometrics that needs to be explained is latent variables. Latent variables are variables that can’t be measured directly, but instead are variables that is based upon multiple other observable variables. A good example of a latent variable can be overall human health as stated in [52]. Here health is a latent (a hidden variable) since it cannot be measured directly but instead depend on multiple other variables such as weight, blood sugar, body temperature, blood pressure etc, which all contributes to human health. Another example from [52] is the measuring of the temperature in a room with four thermometers throughout the day. The temperature will fluctuate depending on the time of day and that can be said to be caused by a latent variable, which is influenced by other variables such as sunlight, air-condition or an open door. The temperature measurements are correlated with this latent variable. Now, the average temperature of the four thermometers can be defined as a latent variable (t_1) by making a linear combination of the measurements and corresponding weights (here each weight is a ¼, since the mean is taken) [52]:

$$t_1 = [\, x_1 \; x_2 \; x_3 \; x_4 \,] \cdot \begin{bmatrix} p_{1,1} \\ p_{2,1} \\ p_{3,1} \\ p_{4,1} \end{bmatrix}$$

The main points of latent variables are that [52]:

- When generated, they can be used to represent the system instead of multiple variables, coming from the raw data, because of the correlation between latent variables and observable variables.
- They capture an unobservable phenomenon in a given system.

Therefore, the use of latent variables is extremely useful when having multiple variables, such in spectrometric data.

4.1 Principal Component Analysis

Latent variables make up the back-bone of the Principal component analysis, PCA, which is used to explain the variance of an observed system. This method includes the generation of a PCA model that consists of two parts, a latent variable model and a residual error.

The principle idea starts in the organisation of data. Data, collected from experiments, can be represented in a matrix, X, in which each row (defined *N*) represent an object. The columns of X (defined *K*) would be the variables, which corresponds to the recorded values for each object. Now, since there is K-variables in X, one can refer to a K-dimensional space, however, for charity, a three-dimensional space is showcased. The following figure, *Figure 12 left*, showcases a three-dimensional space for raw data. [52]

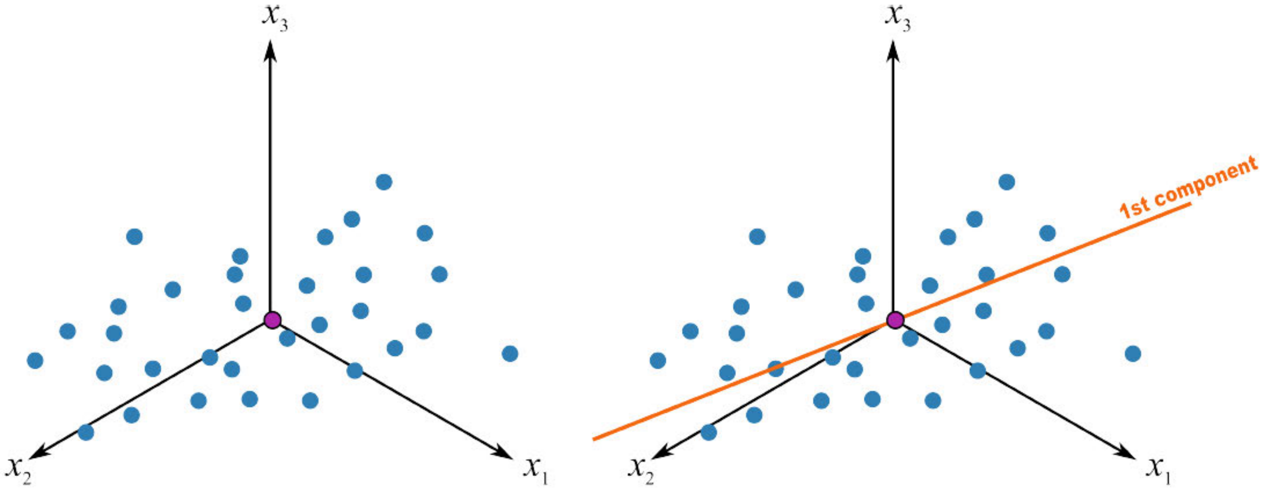


Figure 12: Representation of raw data (left) and first principle component, credits go to Kevin Dunn [52].

Now, firstly the raw data is centred in a data centre as seen in the figure. This is done to remove bias that isn't necessary to model. One can also choose to scale the data to counter the unit variance if the variables have different units of measurement. After scaling and centring, the so-called principal components are generated. The first principal component is fabricated along the direction that gives the biggest variance as seen in the right figure. Now, each object can be projected onto this line by a 90-degree projection. The following figure, *Figure 13*, showcases this in the three-dimensional representation for two objects.

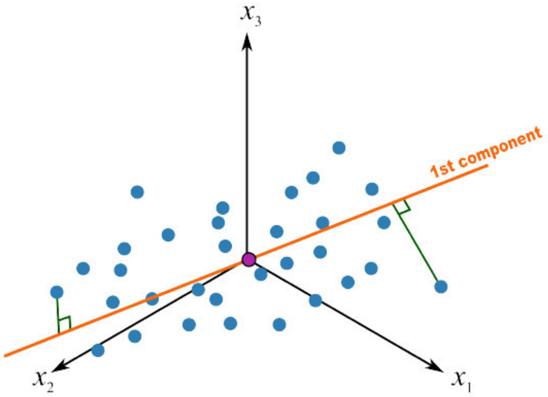


Figure 13: First principle component (PC1) with two objects projected, credits go to Kevin Dunn [52].

An important parameter arises called **scores**, for each object, they can be defined as *"the distance from the origin to the projected point of an object on the principal component (PC) line".* Seen in *cf. Figure 13* this is the distance from origin to every *projected* point. When the PC is drawn based on maximum variance, it is actually based upon the maximal variance of these scores [52]. For charity and on a side-note, variance (s^2) for a sample is calculated the following way:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |\, x_i - \ m|^2$$

Here m is the mean of the sample n.

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The first PC is also the first latent variable, with its structure being made up of a direction vector and the collection of scores along its line. The direction vector is a K-dimensional vector (K x 1) that explains the direction of the first PC in a K-dimensional space, the vector is normally defined as p_1. The magnitude of this vector is irrelevant, since it is rescaled to become a unit vector. The collection of scores is, as seen, also a vector normally defined as t_1 (for PC1) with size N x 1.

The second PC is now found by first generating it perpendicular to the first PC at the start of origin, *cf. Figure 14 (left)*. The second PC can then be rotated around in the space until the greatest variance in the scores is found of this PC when the objects are projected into this new PC line , *cf. Figure 14 (right)*.

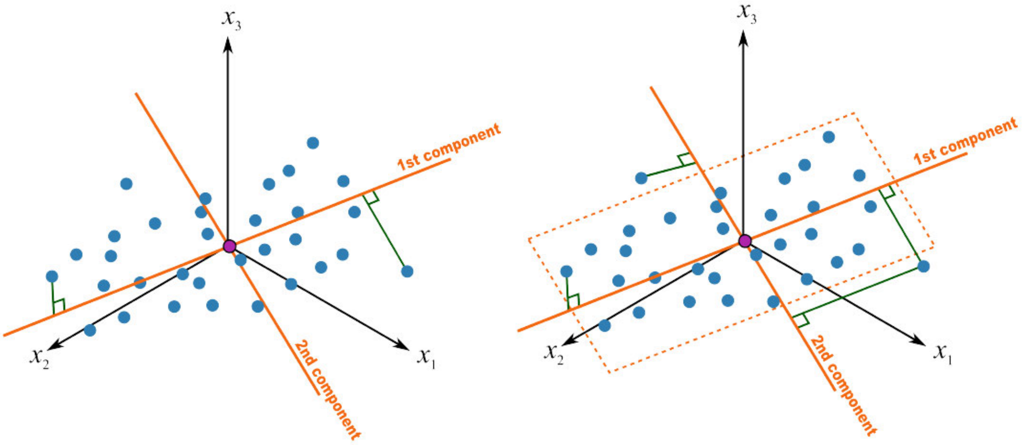


Figure 14: PC1 with two objects projected and PC2 (left). PC1 and PC2, both with two objects projected with the PCA plane (right), credits go to Kevin Dunn [52].

Now, p_2 can be defined as the direction vector of PC2 and t_2 its scores of the projected objects. PC1 and PC2 now spans a plane. This plane is also equal to the latent variable model (containing the p and t vectors) and it can have multiple more components depending if the higher PC explain enough of the variance. One can now define residual error as the perpendicular distance from each object onto the PC plane. When the scores variance is maximized it also follows that the residual error is as small as possible.

Now, as stated previously in the thermostat example, latent variables can be represented by linear combinations and the score values can be represented with the original data and the direction vectors as following [52]:

$$t_{i,1} = x_i' \cdot \mathbf{p_1}$$

Here x_i' is a row from the data (vector size 1 x K) and $\mathbf{p_1}$ is a direction vector (K x 1).

For charity, this translates to [52]:

$$t_{i,1} = x_{i,1} \cdot p_{1,1} + x_{i,2} \cdot p_{2,1} + \cdots + x_{i,k} \cdot p_{k,1} + \cdots x_{i,K} \cdot p_{K,1}$$

$t_{i,1}$ translates to *the score value for Ith object for the first PC*. Remember, that there are K-variables and it is seen on this equation that every variable has its share on the score value (combination of all its ith row of data). The score values for the second component, third and so on follows the same path [52]. The linear combination for the second component as an example is here shown:

$$t_{i,2} = x_{i,1} \cdot p_{1,2} + x_{i,2} \cdot p_{2,2} + \cdots + x_{i,k} \cdot p_{k,2} + \cdots x_{i,K} \cdot p_{K,2}$$

The PCA model in itself have A components and by using a matrix form of these linear combinations all (A) score values can be calculated for a chosen observation [52]:

$$t_i' = x_i' \cdot P$$

Matrix sizes are also included for charity: Matrix P can be explained by saying *there is K rows for all A components*:

$$(1\ x\ A) = (1\ x\ K)(K\ x\ A)$$

Now only scores for one observation has been calculated, however using the whole data matrix one can calculate all scores [52]:

$$\mathbf{T} = \mathbf{XP} \text{ where } \mathbf{P} = [\ \mathbf{p_1}\ \mathbf{p_2} \ldots \mathbf{p_A}\ ]$$

Matrix sizes corresponds to:

$$(N\ x\ A) = (N\ x\ K)(K\ x\ A)$$

This also represents the original data projection to the PC space and explains how one can go from “real dimensions” into PC hyper plane space. The number of A is commonly smaller than K meaning that a big collection of variables can be “stored” in a latent variable (principle component) [52].

On a side note, the real terminology for the direction vectors and direction matrix is loadings.

Data collected from experiments can be said to consist of two parts, a structure, which is the part of the data that can be explained by the latent variables and noise, which cannot be explained by the latent variables. The need arises for two new terms, explained variance and residual (non-explained variance) to help quantify the "randomness" of the experiments. So, PCA provides only a good estimation of a given object using the "collected variables", the latent variables, by projecting a data point to a new plane. The estimates are expressed using the "magnitude" (the score value), and the direction vector (the loadings vector). The matrix notation for the estimates can be reviewed in the appendix of theoretical interest.

Mathematically, the explained variance is calculated by taking the sum of squares of every element in the X(estimate) matrix dividing it with the sum of squares for the whole dataset X as follows:

$$Exp.var. = \frac{sum(X(es)^2)}{sum(X^2)} = Var(X)$$

Residual variance is calculated using the residual matrix (E = X - X(es) = X - TP') in the same procedure:

$$Res.var. = \frac{sum\left(\left(X - X(es)\right)^2\right)}{sum(X^2)} = Var(E)$$

Now, an important factor to evaluate model quality arises, the R^2. The R^2 is defined as:

$$R^2 = 1 - \frac{explained\ variation}{total\ variation}$$

As seen in the definition, R^2 indicates how much of the variation is explained, a R^2 of 1 indicates that the models explain 100% of the variation. This definition can also be applied to the PCA. The residuals can be calculated on a column basis or for the whole matrix. Using the column basis, for the kth column for R^2 the is defined as [52]:

$$R_k^2 = 1 - \frac{Var\left(x_k - x_k(es)\right)}{Var(x_k)} = 1 - \frac{Var(e_k)}{Var(x_k)}$$

More components will increase the $R_k{}^2$ because the estimates are calculated based on the number of components, additional components give additional "correction terms" to the estimate. As said, the maximum is 1, however, it is very unlikely to reach this threshold as the noise also needs to be modelled perfectly. Now, for the whole data matrix [52]:

$$R^2 = 1 - \frac{Var\left(\mathbf{X} - \mathbf{X}(\mathbf{es})\right)}{Var(\mathbf{X})} = 1 - \frac{Var(\mathbf{E})}{Var(\mathbf{X})}$$

For zero components, the R^2 is zero (because no variance is modelled), now each new component explains some amount of variance following this order up to 1 [52]:

$$R_{a=0}^2 < R_{a=1}^2 < R_{a=2}^2 < \cdots < R_{a=A}^2 (= 1.0)$$

R^2 can therefore be said to be a cumulative property for the PCA model.

4.2 Multiple Linear Regression

The use of latent variables can be used in combination with regression to predict a given dependent variable in a method called principal component regression (PCR). It has advantages other its counterpart, multiple linear regression (MLR). However, to see the differences, a brief explanation of MLR will be made.

In multiple linear regression, one has many independent variables and one dependent variable. When having multiple variables, it is necessary to use matrix notation to calculate the weights for each contribution of an independent variable to the dependent one. The dataset X is still a N x K matrix, which can be related to the dependent variable y, a N x 1 vector. The model will take the form y(es) = **X**b, where b is the solution vector. The goal in regression is that the sum of squared errors is minimized, this is accomplished by calculating b, such that:

$$\mathrm{b} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

Now, MLR have several disadvantages when compared to PCR, they can be summarized as stated in source [52]:

- A MLR condition is that the number of objects is larger than the number of variables (N > K). In spectrometric data, this is indeed a problem when one has 100's of wavelengths.

- If the columns in the data X, is strongly correlated, problems can arise. This happens because of multicollinearity, which means that independent variables are dependent on other predictive variables in the model. This problem is seen as; when small changes in the data occurs big fluctuations in the model happens.

- The method cannot handle noise (it models noise) or missing data in X.

Therefore, it is necessary to use other regression methods such as PCR and PLSR. Which have their foundation in PCA and in the latent component space. This provides several advantages compared to MLR [52]:

- The column of scores, that is in the T-matrix, are independent on each other since the PC's are build orthogonal.

- For PCR/PLSR there is no condition that N > K, however, for PCR/PLSR N > A (A is the number of latent variables). This condition is much easier to meet.

- The PCA method in itself separates noise, therefore the T-scores have less noise than X.

The chosen method is PLSR in this thesis therefore a presentation of this method will be presented, however, for further interest in the other method, PCR, a summary is presented in the appendix for further interest.

4.3 Projection to Latent Structures

The method of projection to latent structures (PLS) extends the foundation in PCR. The underlaying principle behind PLS is to have two data sets undergoing PCA, one independent X and one dependent Y. For the generation of the model both datasets are needed, but afterwards one only need to have X to estimate the Y dataset. So, this means that a PLS-model, compared to PCR, can include additional correlated variables in Y and predict them all, instead of predicting one dependent variable at the time as in PCR. Each of the two datasets get their scores extracted and by manipulating these scores, three-objectives of PLS occurs. That is: The PLS model will give the best explanation in the X-space, the Y-space and greatest relationship between X and Y. [52]

Remember, that for PCA, the loadings and scores took the following form:

$$\mathrm{t_a} = \mathbf{X_a}\mathrm{p_a}$$

Making PCA both on the X dataset and Y dataset (as stated in PLS), the following scores and loadings arises for the X-space and Y-space, respectively:

$$\mathrm{t_a} = \mathbf{X_a}\mathrm{w_a}$$

$$\mathrm{u_a} = \mathbf{Y}_{\boldsymbol{a}}\mathrm{c_a}$$

The loadings $\mathrm{w_a}$ and $\mathrm{c_a}$ are also constrained to unit length.

These score vectors for both spaces need to have maximal covariance in order to achieve the three objectives. This will be further elaborated.

Firstly, a function for mean, variance, covariance of a vector and correlation will be defined to make the formulas more compact. They are respectively:

$$\varepsilon\{x\} = m = \frac{1}{n}\sum\nolimits_n x \text{ (definition of mean)}$$

$$V\{x\} = \varepsilon\{(x-m)^2\} = s^2 = \frac{1}{n}\sum(x-m)^2 \text{ (definition of variance)}$$

$$V\{x\} = \varepsilon\{(\mathrm{X} - \varepsilon\{\mathrm{X}\})(\mathrm{X} - \varepsilon\{\mathrm{X}\})^{\mathrm{T}}\} \text{ (definition of covariance of a vector X)}$$

$$\mathrm{Cor} = \frac{\varepsilon\{(x-\bar{x})(\mathrm{y}-\bar{y})\}}{\sqrt{V\{x\}V\{y\}}} \text{ (definition of correlation)}$$

Now covariance can be defined as the following between the two score vectors for each space:

$$\mathrm{Cov}(\mathrm{t_a}, \mathrm{u_a}) = \varepsilon\{(\mathrm{t_a} - \overline{\mathrm{t_a}})(\mathrm{u_a} - \overline{\mathrm{u_a}})\}$$

Covariance defined in terms of correlation takes the following form:

$$\mathrm{Cov}(\mathrm{t_a}, \mathrm{u_a}) = \mathrm{Cor}(\mathrm{t_a}, \mathrm{u_a})\ \mathrm{x}\ \sqrt{V(\mathrm{t_a})}\ \mathrm{x}\ \sqrt{V(\mathrm{u_a})}$$

$$\mathrm{Cov}(\mathrm{t_a}, \mathrm{u_a}) = \mathrm{Cor}(\mathrm{t_a}, \mathrm{u_a})\ \mathrm{x}\ \sqrt{\mathrm{t'_a t_a}}\ \mathrm{x}\ \sqrt{\mathrm{u'_a u_a}}$$

If one should give an interpretation of covariance, the higher the number, the greater the relationship between two vectors, correlation goes under the same principle, but its values only ranges between -1 to 1 (for strong negative or strong positive correlation). The mean of the score vectors is zero, therefore, the variance term can be rewritten in terms of a dot product (here the definition of covariance for a vector is used). The covariance then comes down to maximizing the dot product between the two vectors t_a' and u_a.

The three objectives occur from the above-mentioned formulae statement:

- The best explanation of the X-space by having the t_a't_a part.
- The best explanation of the Y-space by having the u_a'u_a part.
- The greatest relationship between XY by having the correlation part between the scores.

Therefore, the objective in PLS is to have maximum covariance. The algorithm for PLS modelling has been excluded from the master thesis to not get side-tracked.

4.4 Cross-validation

Cross-validation is a statistical tool that can be used on any model to avoid over-fitting and is therefore an important aspect for validation of models. For models based on latent variables every addition of a new latent variable will explain the X data more, so the R^2 will continue to increase. Now, the problem with this is that at a certain number of latent variables, the variables will extract all systematic variation out of the X data. Continuing adding additional latent variables will model the noise in the dataset and give the illusion of a very good predictive model, the model would now be overfitted [52]. Normally, one could make a new dataset (a test set) and see how the current model predicts this new set, if there is bad prediction, overfitting has occurred. However, in many cases it is not possible to make a test set, timewise or economically. A way out of this problem is to use the current dataset and the methods of cross-validation to combat overfitting.

The principle will now be discussed. First, the dataset X is divided into groups *g* on a row-basis. The selection can be done in multiple different ways: randomly, leave-one-out, group-order-wise. The groups are collected into a new dataset X_1, the remaining dataset can be defined as X_{-1}. In general, for each *g* group one can generate X_g and X_{-g} datasets, where the amount goes from [1…i…g]. Now, a model is made using the X_{-i} datasets (the remainder data). The X_i (excluded values) are used as a testing set to calculate new y-values for a dependent variable. This procedure can be repeated so all X values have been excluded and new y-values estimated. For every excluded x value, one can then calculate the residuals E_1…E_i…E_g on a step-wise leave-out-data model generation basis:

$$\mathrm{E_i} = \mathrm{X_i} - \mathrm{X_i}(es)$$

A new R_{cv}^2 can now be defined for the A number component:

$$R_{cv}^2 = 1 - \frac{Var(E_{A,cv})}{Var(X)}$$

To summarize the method of cross-validation the basics are:

1. The dataset X is divided into groups, g.
2. For each step from i to g:

- Ith group is excluded from the original set.
- The model is calibrated based on the modified set.
- The model now predicts y-values for the excluded group g.

So, a set of both cross-validation and normal statistics are available such as root-mean-squared-error-of-prediction (RMSEP) and R^2. It is interpretation of the two statistic sets that determine of a model overfits or not, where R$^2_{CV}$ and RMSEP$_{CV}$ is particularly useful. This will be discussed. The definition of the RMSEP is the following, the hat indicates an estimate:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

For a model to have a very good predictive nature, this value should be as low as possible.

R$^2_{CV}$ behaves as R^2 and explains how much of the variation is explained in the cross-validated set. Two important properties of R$^2_{CV}$ is first, that it is for the most part smaller than R^2 and second, at some number of latent variables the R$^2_{CV}$ doesn't increase with new latent variables added to the model. The value will instead tend to decrease giving a very good indication of where the optimal number of components are located. This is because when the new latent variables are introduced, they can't explain anymore systematic variation in the cross-validated dataset. The properties for RMSE behaves similarly.

Therefore, an important step in choosing a good model is either looking at R^2 vs. R$^2_{CV}$ or RMSEP vs. RMSEP$_{CV}$. The best models are normally found when the cross-validated statistics are at a maximum for R^2 or minimum for RMSE. An example from [52] illustrates this point, *cf. Figure 15* and showcases when the number of components are optimal.

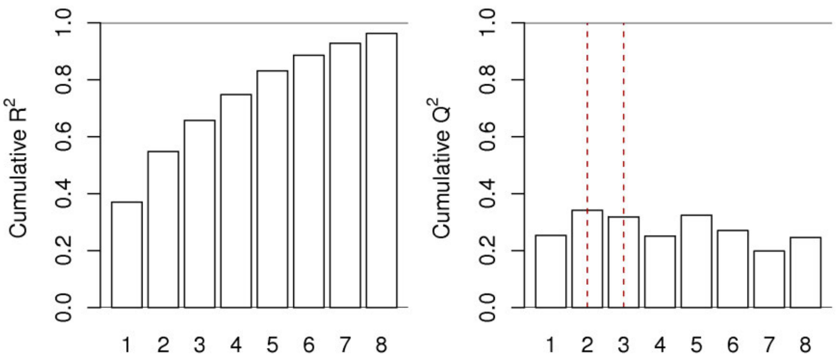


Figure 15: Cumulative plot of R2 (left) and R$^2_{CV}$ (Q^2 right) vs. latent variables. Credits go to Kevin Dunn [52]*. Picture under creative commons licence.*

Now, finding the best number of components can be somewhat subjective. Here the optimal number of components can either be two or three components depending if the 3rd showcase interesting changes.

The fifth component should be avoided because of its possibility to model noise. This figure also illustrates the importance of cross-validation since the R^2 just tends to increase and thereby overfit more and more.

Now in this master thesis, the optimal number of components are expected to be found this way when R^2_{CV} is at the first maximum or $RMSE_{CV}$ is at the first minimum. However, there exist another method for determine the optimal number of components. This will be reviewed shortly.

4.5 Wold’s R

There exists another method for finding the most optimal number of components instead of the traditional method of minimum RMSE. This method takes its point of origin in the values, prediction sum of squares (PRESS) [53]. The PRESS values are defined as:

$$\mathrm{PRESS} = \sum_{i=1}^{N} \left(y_i - \widehat{y_{\imath(\imath)}}\right)^2$$

Now, the second term are the fitted values from cross-validation. The hat means that it’s an estimate of y, the first subscript means it’s for the i^{th} datapoint, the second i^{th} means that it has been estimated on a model with the i^{th} estimate excluded. It is seen, that smaller PRESS values mean better models and should, based on the number of components, encounter a minimum and hereafter rise again. On that note, the study now defined a Wold’s R value that are dependent on the number of components, a [53]:

$$\mathrm{R} = \frac{\mathrm{PRESS(a+1)}}{\mathrm{PRESS(a)}}$$

So, it is basically a ratio between the PRESS values of two latent variable, with the succeeding latent component in the numerator. When the R-value is under unity, it means that the succeeding latent component has lower PRESS value and it would be advantageous to include it. Therefore, when incorporating this in model generation, the threshold for optimal number of components should activate when this threshold passes unity. The optimal number of components would therefore be *a* [53].

However, as stated in the study [53], this criterion seems not to give desirable results when the data matrix got more predictor variables. The spectrometric data in this master thesis have many predictor variables, therefore the method would not be the focus when finding the optimal number of components but will be tried alongside other chosen methods.

4.6 Pre-processing and Variable Selection

One aspect that also has been an important part of chemometrics is the use of pre-processing and variable selection methods. The main function of these methods is to improve the pillars of chemometrics: the exploratory analysis, the multivariate regression and classification models [54]. The need for pre-processing arises because of systematic variations that can emerge in solid samples. In solid samples difference in effective path length and the way that light scatters in its micro-structure gives these variations, that can be seen as baseline shifts or non-linearities in chemical spectra. Non-linearities can be spotted if there is curvature in a predicted vs. measured plot of a given variable [54]

There exist two subclasses of pre-processing techniques: scatter-correction techniques, which include multiplicative scatter correction (MSC), standard normal variate (SNV) and normalization. As the name suggests, these techniques account for the variation due to scatter and hereby baseline shifts seen in ex. the NIR spectra. MSC also accounts for multiplicative effects. The second subclass is Spectral derivative techniques such as Savitzky-Golay (SG) polynomial derivative filters. This subclass removes additive (baseline shifts) and multiplicative effects (sloping of spectra baseline) in chemical spectra [54]. These are the most common pre-processing techniques and this review will only focus on them, even though there exist many other.

On a quick note, one could also choose to pre-process data using mean centring (subtraction of the mean of a variable for the objects), standardization (subtraction of the standard deviation from each object with the corresponding variable) or both in an autoscaling procedure. These methods help interpretation of the data ex. When the data have different units of measurement, then standardization is useful, since its helps comparison between variables of different units. These methods also keep any systematic variation in the data. For a given data matrix X with i rows and j columns, the formulas for mean centring, standardization and autoscaling respectively correspond to:

$$x'_{ij} = x_{ij} - \overline{x}_j \,,\ \overline{x}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij} \text{ (mean centring)}$$

$$x'_{ij} = \frac{x_{ij}}{s_j} \,,\ s_j = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\left(x_{ij} - \overline{x}_j\right)^2} \text{(standardization)}$$

$$x'_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \text{ (autoscaling)}$$

4.6.1 Standard Normal Variate and Normalization

There is much similarity between mean centring/standardization and standard normal variate (SNV) and normalization. This is best explained by showcasing the formulas for both methods, the top formulae is SNV and the bottom one is normalization.

$$x'_{ij} = \frac{x_{ij}-\bar{x}_\imath}{s_i} \text{ (SNV)}$$

$$x'_{ij} = \frac{x_{ij}}{s_i} \text{ (normalization)}$$

As one can see, the SNV functions on a row basis (i) instead of a column basis (j) as in autoscaling. So, this means (for chemical spectra as data) that SNV subtracts the mean of a chemical spectrum for each element in it and then divide by the standard deviation of that given spectrum. This method will eliminate much of the baseline shift.

4.6.2 Multiplicative Scatter Correction

Now multiplicative scatter correction (MSC) can account for both scatter effects (sloping and baseline shift) and has some similarities with SNV. It takes focus in ordinary linear regression and it consists of two parts:

1. The method starts to estimate the contributions of the additive (A) and multiplicative (B) effects. This is done by making a linear model between the original spectrum (x_{org}) and a reference spectrum (x_{ref}). Normally the reference spectrum is just the mean spectrum of the dataset. Mathematically, this is formulated as:

   $$x_{org} = A + B \cdot x_{ref} + e$$

   With linear regression, A and B are estimated for each object.

2. Now, one can correct the spectrums by using the estimates in a likewise way as in SNV. Subtraction of A removes the baseline offset and division with B removes the spectral slope.

   $$X'(i,k) = \frac{X(i,k) - A(i)}{B(i)}$$

   The dataset should now be corrected.

4.6.3 Savitzky-Golay Polynomial Derivative Filters

This method works a little differently compared to the previous methods. Firstly, this method includes a smoothing step of the data, this is done by a piecewise polynomial approximation of the datapoints. This is done because the method needs to find derivates to the polynomial function. [54]

In a chosen window for each point of the spectrum, the polynomial is fitted and the parameters for the estimated polynomial are found. The smoothed signals are used instead of the original signals. Now, for each datapoint, the derivate can be found for any order. However, normally it is only the first and the second order derivates that are useful. Lastly, it is now these derivate values that are used instead of the original values thus eliminating scatter effects (slope and baseline shift). The first derivate effectively removes the effects of the slope and the second deviate removes the effects of the baseline shift, *cf. Figure 16*. [54]

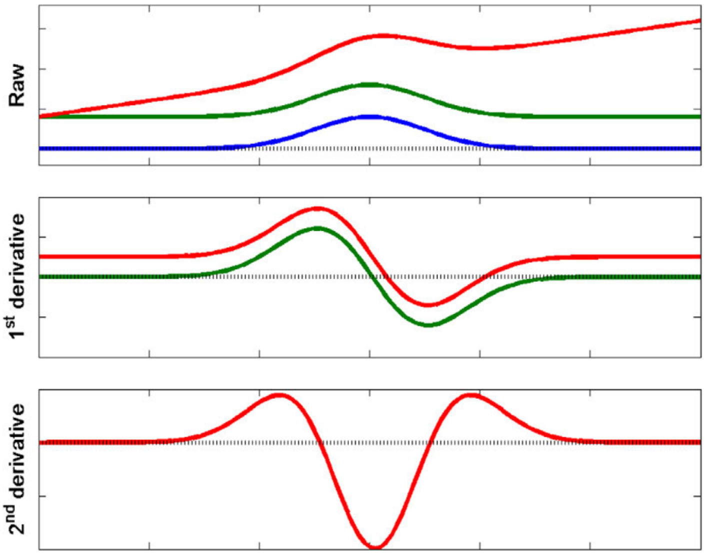


Figure 16: Effect of the savitzky-Golay derivative polynomial filter. Red is a spectrum with both multiplicative and additive effects. Green only has additive effect. Permission for image granted by the source [54].

In this method it is therefore necessary to evaluate a) what derivative order to look at, b) the size of the smoothing filter and c) what polynomial degree to use for approximation.

4.6.4 Model-based Variable Importance Methods

Variable selection methods are methods which reduces the number of variables, by removing noisy and irrelevant variables, and selecting the most important ones for model generation. This has several advantages such as minimizing overfitting, reducing model complexity, hereby making better interpretation of the model and lowering measuring cost. The problem of overfitting occurs if the number of objectives are much smaller than the number of variables, in that case, the risk to find a *random* relationship between the dependent and independent variables are greatly enhanced [55].

One method is called Variable Importance in Projection (VIP), this is a measure of how much a variable explains ("describes") the two datasets X and Y. A value greater/close to 1 is very important for the model and variables with smaller numbers are less important. The VIP score is calculated as following [55]:

$$VIP_j = \sqrt{\frac{\sum_{i=1}^{A} w_{ji}^2 \cdot SSY_i \cdot N}{SSY_{tot} \cdot A}}$$

The symbols are defined as following:

- N is the number of variables
- A is the number of components
- W_{ji} is the weights for variable j and component i. Weights have not been explained, but one can see them as the precursors of the loadings. In the PLS algorithm for calculation of scores, the weights are a set of regression coefficients used.[52]
- SSY_i is the sum of squares of explained Y variance for the component of i
- SSY_{tot} is the total sum of squares of explained Y variance.

Another method is called selectivity ratio (SR) and this ratio is calculated between explained variance and residual variance for each variable. Small SR values corresponds to less important variables.

Since both methods calculates variable importance on single variable basis, it is very easy to find the most important variables by plotting them and removing variables under a given threshold set by inspection of the VIP/SR plots. It is stated in [55], that removing all variables under 1 for VIP is not preferable and caution most be advised.

Another model-based variable selection method is jack-knifing, which is more computational heavy then the two other methods [55]. To use Jack-knifing, one must use the models generated doing cross-validation and their statistic estimates. Jack-knifing will then calculate an uncertainty estimate for each of these statistics and determine ex. of a regression coefficient is significantly different from zero. From this, the least significant variables can be excluded, and one can then choose to repeat the procedure to remove additional variables.

The uncertainty estimate is calculated from:

$$\text{se} = \sqrt{\frac{n-1}{n}\sum (p(i) - \bar{p})^2}$$

Here se is the uncertainty estimate, n is the number of models used, i indicates the given cross validated model, p is the chosen parameter, the line above p indicates the mean for all models.

The other class of variable selection methods (methods based on model comparison) includes making different models each with a different subset of variables and comparing them together. These methods tend to be more computational heavy. One that is primary used for spectroscopic data is the “interval PLS” (iPLS) method. This method looks at sets (intervals) of variables instead of individual variables, this is useful for spectroscopic data since the variables tend to be strongly correlated [55].

The method works by dividing the data into equally sized variable intervals followed by generation of a set of PLS models, one for each interval. Interpretation becomes much easier since one can look at which models gives the best predictions compared to others. Therefore choosing the right interval size becomes of importance [55]. With the models generated one can select which variable groups gives the best model and combine them one by one in a “forward iPLS” approach or make the full model and remove the variable groups which gives the worst models one by one in a “backward iPLS” approach [55].

A table of other non-commonly methods are showcased here, *cf. Table 4*, but are not evaluated further to limit the scope of variable selection methods in this master thesis. This is not a complete list and there exists many others.

Table 4: Other variable selection methods. [55][56]

Variable Selection method	Description
Generic algorithms (GA)	Producing of models based on an evolution generation approach. It takes different combinations of variables and makes a model on them. The model then evolves through several model generations by adjustment of the number of variables.
Least absolute shrinkage and selection operator (LASSO)	Here a threshold is chosen based on the absolute size of a given regression coefficient. If the coefficient doesn’t uphold this threshold, the coefficient will be reduced to zero in an automatic approach and the corresponding variable will be removed.
Competitive adaptive reweighted sampling (CARS)	Very advanced variable selection procedure. Based on multiple advanced variable selection and sampling procedures the variables “compete” so only key variables are extracted for a final model. The study [56] compared this method will Darwin’s evolution theory “survival of the fittest” (variable).

5 Preliminary Experiments

The following sub-chapters are made to explain and discuss the prediction procedures and reference methods for the three quality parameters chosen: protein content, relative hardness and vitreousness together with an introduction to the wheat samples.

5.1 Introduction to wheat samples

The ten wheat samples provided from the project partner CSORT are all spring wheat that are specifically breed for western Siberia. They were grown on an experimental field of the Altai Research Institute of Agriculture (ARIA) with harvest in 2015. 7 of them are classified as soft wheat and 3 of them classified as hard wheat. The full names of the wheat samples are as following:

- Altayskaya 75 (Alt75)
- Altayskaya 105 (Alt105)
- Altayskaya 325 (Alt325)
- Altayskaya Stepnaya (Altstep)
- Altayskaya Zhnitsa (Altzh)
- Altayskaya 530 (Alt530)
- Sibirskiy Alians (Sibirs)
- OASIS (OASIS)
- Salyut Altaya (Salut)
- Pamyati Yanchenko (Pam)

When referring to these wheat types, the short names in the parentheses will be used. Three different protein determining tests have been conducted on the wheat sorts, *cf. Table 5*.

Table 5: Protein content of the wheat sorts with hardness class.

	Hardness Class	Protein [1 - %]	Protein [2 - %]	Protein [3 - %]
Alt75	Soft	15.45	NA	NA
Alt105	Soft	15.20	NA	NA
Alt325	Soft	16,55	NA	14.12
Altstep	Soft	14.15	NA	NA
Altzh	Soft	14.40	NA	14.76
Alt530	Soft	15.00	NA	NA
Sibirs	Soft	16.60	NA	NA
OASIS	Hard	NA	11.00	10.80
Slaut	Hard	NA	14.28	14.90
Pam	Hard	NA	12.21	NA

The first protein test was conducted in Jan 2016 by the official laboratory of Federal Center of Grain Safety and Quality with the Kjeldahl method. The second test was conducted in Mar 2017 by the same institute and the same method. The third test was conducted in 2018 as part of a Bachelor Thesis at Altai State University, also with the Kjeldahl method. In this thesis the results from the first and second test will be used for comparison purposes.

5.2 Vitreousness Determination

Unfortunately, it was not possible to get the right equipment to evaluate the quality parameter, vitreousness and it was thought to exclude this parameter all together because of its subjective nature of determination. However, shortly before running the main experiments, the vitreousness was nonetheless included in this thesis and evaluated by using sub-optimal equipment such as a pill-cutter and microscope instead of the special cutting device, the farinator. The procedure for measuring vitreousness was a modified version of the standard method of visual evaluation. It was conducted as follows:

1. The kernels would be placed in the pill-cutter and quickly sliced to get the clearest cut.
2. Afterwards, the kernels were visually inspected using a standard microscope. In the main experiments, damage on the kernel would sometimes occur making it hard to distinguish which of the cross-section were vitreous and which seemed vitreous because of the factures in the kernel.
3. The kernels were tried classified into three groups, non-vitreous (A), piebald (B) and fully- vitreous (C).
4. Together with classification a guess was done on how much % of the cross-section seemed vitreous.

The classing of the virtuousness was done according to the classing in [40], *cf. Figure 17*.

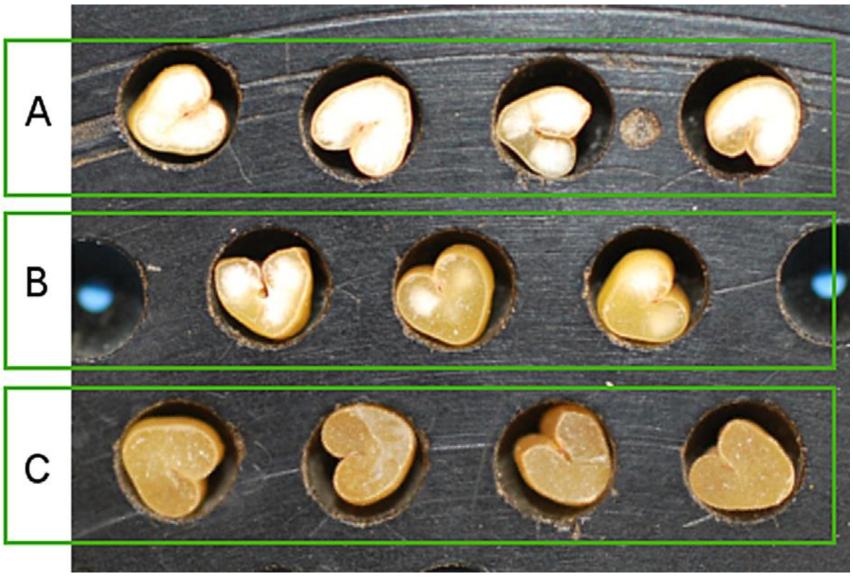


Figure 17: Different classes of vitreous kernels. A is non-vitreous, B is piebald and C is fully vitreous. Permission to use this image has been granted by the source [40].

30 kernels were visually evaluated to get experience with the parameter otherwise no preliminary experiments with vitreousness was conducted other than around +9 kernels that were evaluated but data was not recorded for them. The data for the 30 kernels can be showcased in the appendix and consist of three kernels from each of the ten sorts. The interesting part about the evaluation was that many kernels seem to be fully vitreous. The pill cutter also gave damage to the kernels making it hard to see which class each kernel belonged to. The following chapters will be dedicated to relative hardness and protein content since no clear analysis can be done on this small amount of data.

5.3 Protein Determination with FIA

The main objective of the preliminary experiments with FIA was to determine if the method could accurately predict the protein content in whole single wheat kernels from different sorts. A criterion for success would be if the distribution of the protein in the wheat follows the same pattern as the protein content found in the method of Kjeldahl provided by the partner Csort LTD.

The procedure that was followed was application note 5202 from FOSS *“Determination of total oxidized nitrogen in water by FIASTAR 5000”.* The description in this application note states that this method can determine nitrogen present in water. The species include ammonia, ammonium, nitrate, nitrate and organic nitrogen compounds that can undergo conversion to nitrate under oxidative conditions. The calibration range goes from 0.1 - 5 mg/L N. The calibration is linear.

Based on the calibration range, it was necessary to evaluate the limits of the calibration curve to not reach the upper limit. A study suggested that the protein content in wheat kernels ranged from 6.15% to 19.77 % with a mean of 10.59% and standard deviation of 2.07% [14]. Therefore, the maximum threshold of a kernel’s protein content was set to 20%. With the updated protein conversion factor of 5.83, the upper threshold of the maximum weight of nitrogen (WN) for a upper threshold of 5 mg/L N in a 50 ml volumetric flask was determined as:

$$\mathrm{WN} = 5.83 \cdot 5\,\frac{\mathrm{mg}}{\mathrm{L}} \cdot 0.050\ \mathrm{L} = 1.45\ \mathrm{mg}$$

Now based on the maximum nitrogen amount and the upper limit of 20%, one can calculate the max weight of a given (WS) sample:

$$20\% = \frac{1.45}{\mathrm{WS}}$$

$$\mathrm{WS} = 7.25\ \mathrm{mg}$$

Therefore, if a sample is less than 7.25 mg, the nitrogen amount should be inside the calibration range.

Now, there are several chemicals that are used in the application note, the whole list can be reviewed in the appendix. However, the main principle behind this procedure and generation of the most important solutions are as following:

5.3.1 Procedure for Digestion

This is the specific procedure followed when digesting wheat kernels.

1. Chosen kernels for analysis are pulverized by mortar and pestle. Three representative samples of each unique kernel flour were taken, and each flour sample was weighed carefully.

2. Each flour subsample was moved to a 100 ml autoclave flask with a unique ID (ex. 1.3 for kernel 1 flour 3). The weighing boats used for subsampling was all sprayed with distilled water to capture the leftover when transporting the flour to the autoclave flasks.

3. The digestion solution was then added to the autoclave flasks.
   Here, the alternative digestion procedure was followed. As stated previously, the alternative digestion includes oxidation of the samples using peroxydisulfate.

   The digestion solution was made by adding 25 g of peroxodisulphate ($K_2S_2O_8$) and 15 g of boric acid (H_3BO_3) in a 500 ml hydroxide solution (0.375 M).

   To each autoclave flask, 8 ml digestion solution was added

   To each autoclave flask, 40 ml distilled water was also added.

4. Each flask was autoclaved for 30 minutes at 120 C°. The samples were cooled at room temperature overnight.

5. Optimally, now the wheat flours nitrogen should all be oxidized to the form of nitrate.

5.3.2 Procedure for FIA after Digestion

1. Afterwards all the contents of the autoclaved flasks were now filtered, by ordinary paper filter, to a volumetric flask of 50 ml. The filter paper was rinsed with distilled water. The volumetric flasks were afterwards diluted to the 50 ml mark.

2. After filtration, a sample was taken from each volumetric flask and placed in a sample tube in the FIA sampler. A 2 mg/L glycine validation was likewise added.

3. These are now mixed with a buffer solution in the FIA.

   The buffer solution was made by dissolving 85 g ammonium chloride NH_4Cl, first in 500 ml distilled water in a 1000 ml volumetric flask. The pH was then adjusted to 8.5 by using common bases (NH_4OH) and acids (HCl). Then the contents were diluted to 1000 ml.

4. Afterwards, a cadmium reductor reduces the nitrate to nitrite.

5. By adding an acidic sulphanilamide solution, the generated nitrite will form a diazo compound.

   The sulphanilamide solution was made in a 500 ml volumetric flask. Firstly, 5 g sulphanilamide was diluted in 250 ml distilled water, afterwards 25 ml concentrated hydrochloric acid was added. Finally, the solution was diluted to the mark on the volumetric flask.

6. Now, the diazo compound is combined with NED regens (N-(1-naphtyl)-Ethylene Diamine Dihydrochloride) to produce a purple azo dye. This purple dye can be measured at a wavelength of 540 nm.

   The NED solution was made in a 500 ml volumetric flask, firstly 0.5 g NED reagent was diluted in 250 ml distilled water. Afterwards, the volumetric flask was filled to its mark.

5.3.3 Equipment and Verification

Since it is possible for the cadmium reductor to expire, it was necessary to check its reduction efficiency. This was done by injecting a 1 mg/l NO_2-N (nitrite) solution over the reductor, afterwards, a 1 mg/l NO_3-N (nitrate) was injected. The NO_3 injection should correspond to the NO_2 value if the cadimium reductor is running at 100% efficiency. This was not the case and therefore a completely new cadmium reductor was introduced before running any preliminary experiments.

The efficiency of the cadmium reductor was checked once more before running of the main experiments, the efficiency was satisfying ~98%.

The stock solution used for calibration was a 20 mg/l NO_3-N (made from a pre-made 1000 mg/l NO_3-N). The working standards were made as following:

Table 6: Standards for calibration

NO_3-N concentration (mg/l)	Volume of 20 mg/l (interim standard)	Final volume (mL)
0	-	100
0.1	1	200
0.5	2.5	100
1	5	100
3	15	100
5	25	100

The calibration solutions were made freshly for every batch run.

To ensure that the calibration was done correctly, a 5 mg/l NO_3 was run after each calibration together with a blank. Furthermore, after each batch of samples a 5 mg/l and a blank were included to test if the calibration was still valid. If the batch number was larger than ten samples, then for each tenth sample a quick calibration with a 5 mg/l was conducted to ensure stability.

For verification of the digestion a 2 mg/l glycine solution were digested together with each kernel batch. The 2 mg/l glycine solution was made fresh weekly from a 200 mg/l stock solution (stock made once a month). The glycine sample was 40 ml 2 mg/l glycine and 8 ml digestion solution. So, the real concentration of the glycine solution was instead 1.66 mg/l. The criterion for successful digestion was therefore that the concentration of glycine didn't come under this new concentration.

However, glycine is not a complete candidate "to mimic" the complex matrix of wheat kernels and the comparison between digestion of glycine and digestion of wheat kernels may therefore be flawed.

Possible errors that arise doing protein determination with the alternative digestion procedure and FIA could arise from weighing error, digestion error (lack of proper digestion), contamination, calibration and in the creation of custom reagents. Especially temperature seems to influence the preliminary experiments. A study suggests that the errors are significant in regard to time and temperature [28].

The application note specifically states that variations happens based on temperature, the condition of the tubing's and the purity of reagents. To verify the accuracy of the overall ability to predict protein with this method, reference protein contents have been delivered. However, these reference results are based on the Kjeldahl method on a bulk basis and variations may therefore arise.

5.3.4 FIA: Sampling and Results of Preliminary Experiments - Part I

First, three kernels were sampled randomly from one of the ten sorts, OASIS and analysed following the procedure described previously. This was done in order to get experience with doing the experiments ex. making the solutions, calibrating the equipment. No glycine validation was performed on these.

To differentiate the kernel data, each kernel got a number ID and each powder sample got a letter ID. An example is showcased for the weights and the found nitrogen content of these three samples in *cf. Table 7.*

Table 7: Weight and nitrogen content table for the three OASIS kernels.

	Weight [mg]		
	A	B	C
1	10.1	7.1	10.1
2	6.5	7.6	5.7
3	8	9.5	7.4
	Nitrogen content [mg/L]		
	A	B	C
1	2.778	1.802	2.691
2	1.540	1.807	1.231
3	1.852	1.869	1.637

When analysing the nitrogen content, the powder samples were not randomized. However, all main experiments got their order of nitrogen measurements randomized.

Now, a semi-automatic calculation procedure was developed in Excel to calculate the protein content. Excel was chosen because of its quick overview to ensure every data was placed and written correctly.

An example for nitrogen to protein conversion can be presented for A1, see *Table 7*:

1. It is assumed that all nitrogen is trapped in the 50 ml volumetric flask containing the wheat kernel powder leftover. Therefore, a conversion from mg/L to mg/50 ml is done:

   Traditional dilution formulae:

   $$\frac{V_{end}}{V_{start}} = \frac{C_{end}}{C_{start}}$$

   Numbers placed, C_{start} isolated and calculated:

   $$C_{start} = \frac{C_{end}}{\frac{V_{end}}{V_{start}}} = \frac{2.778\,\frac{\text{mg}}{\text{L}}}{\frac{1\ \text{L}}{0.050\ \text{L}}} = 0.1389\frac{mg}{50\ ml}$$

2. Afterwards the amount of protein is calculated using the conversion factor of 5.83:

   $$\mathrm{C}_{\textbf{protein}} = 0.1389\frac{mg}{50\ ml} \cdot 5.83 = 0.809\frac{mg}{50\ ml}$$

   Since it is assumed all nitrogen from the sample is trapped in 50 ml, then the sample would contain 0.809 mg of protein.

3. Using the weight of the kernel powder, *cf. Table 7*, the protein content on a (%) mass basis can now be calculated:

   $$\text{w}(\%) = \frac{0.809\ \text{mg}}{10.1\ \text{mg}} \cdot 100 = 8.01$$

This calculation is done for all current and future samples and the excel dataset will produce the following: Protein content (%) for each powder sample *cf. Table 8*, mean and standard deviation for each kernel *cf. Table 9*. The mean is used as the cumulative feature used in model building and the standard deviation is used to observe any faulty powder samples.

Table 8: Protein content (%) in each powder sample.

	A	**B**	**C**
1	8.017	7.398	7.766
2	6.906	6.930	6.295
3	6.748	5.734	6.448

Table 9: Mean and standard deviation for the three kernels.

Kernel Nr.	Mean	Standard deviation
1	7.727	0.254
2	6.710	0.293
3	6.310	0.425
Overall	6.916	0.596

Henceforward, the data for this kind of experiments will be placed in the appendix or as an excel file attachment and only critical data or noteworthy data will be showcased in the main thesis. Already now, it is seen that the protein content seems on the low side. The method of Kjeldahl for OASIS wheat showed 11 Protein (%) whereas here it is ~7 (%). It is therefore of upmost importance to examine if this difference is random or systematic in nature before measuring anything used to model building. Therefore, a bigger initial experiment was conducted to test the randomness and stability of this method.

5.3.5 FIA: Sampling and Results of Preliminary Experiments - Part II

In this part two sorts were chosen Alt75 and OASIS. The sorts were chosen based on two criteria 1) the difference in appearance, *cf. Figure 18*, A75 seems a lot darker than OASIS and 2) the difference in protein content using the Kjeldahl method was high 15.45 % for A75 and 11 % for OASIS. If the experiments using FIA could also showcase this difference in a likewise manner, it could mean there is a systematic variation between the methods. 12 random kernels were chosen from A75 and 12 chosen random from OASIS.

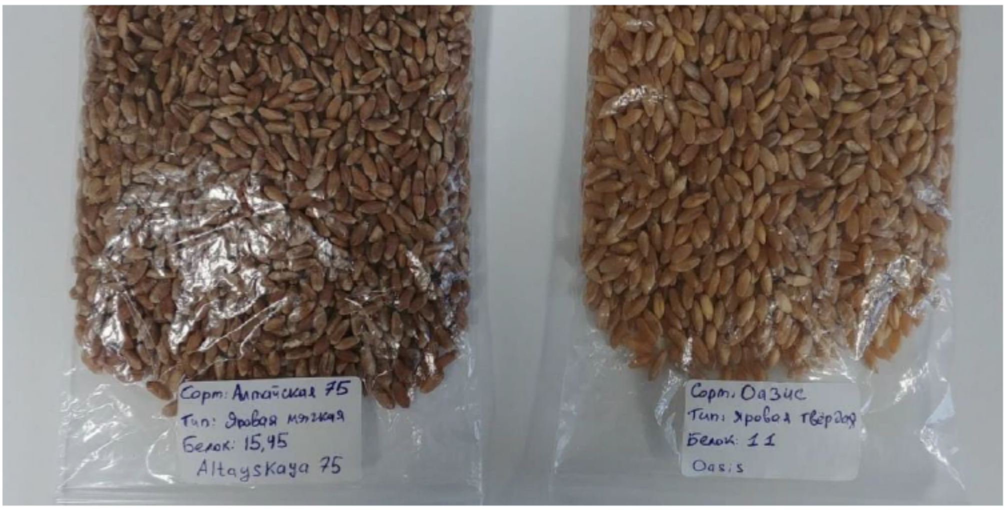


Figure 18: Visual comparison between Alt75 (left) and OASIS (right).

One additional property that is tested for are if there is any difference between the powder samples. It was observed in the initial trail experiments that large flaks of the kernels outer layer, the bran, were produced under grinding. When sampling the powders these flaks would normally enter the first powder sample. It would be interesting to see if this caused any significant effect between the powder samples.

The data for these experiments are all placed in the appendix. If the standard deviaiton of a given kernels powders reaches 1 or over then the anomaly powder in that kernel is deemed outlier. Outliers are marked in red and are therefore not used. The mean is then based on two powders instead of the three.

To summarize: each power sample can be split into two groups, main group one (X1) consists of the three powder samples, subgroup 1 for the first taken, subgroup 2 for the second taken and subgroup 3 for the third taken sample of each kernel powders. Group two (X2) are the sort, OASIS is subgroup 1 and Alt75 is subgroup 0.

To analyse if there is any difference between the group means of multiple groups, n-way analysis of variance (ANOVA-N) has been conducted on the mean protein content. The calculated p-value for group one (X1) is 0.471 which indicates that the means of the powders are not significantly different. It is hereby confirmed that there is no significant difference between the first sampled and the last sampled powders. The p-value for X2 is 0 indicating that there is significant difference between the means of the two wheat samples. The ANAVO-N results is illustrated below, *cf. Figure 19*, with the use of Tukey’s multiple comparison test.

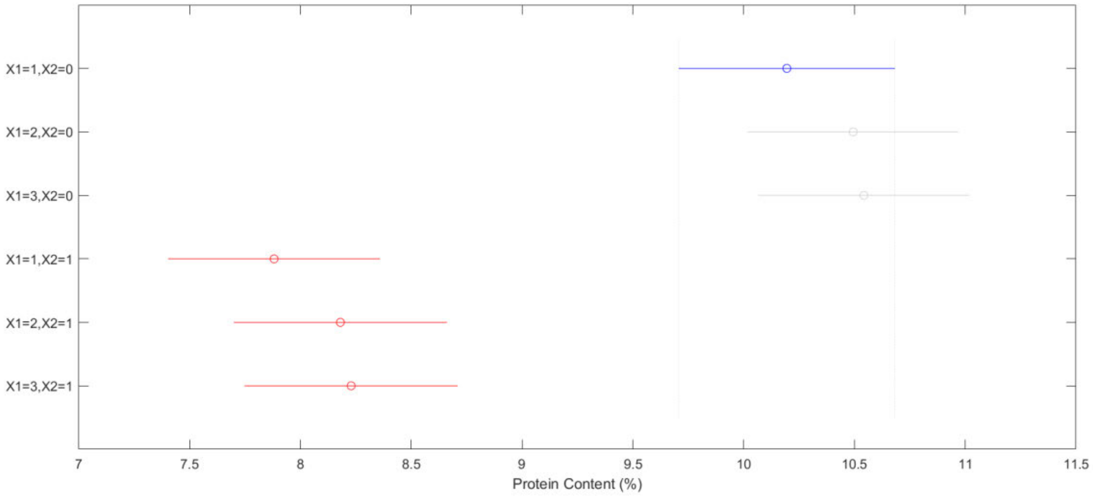


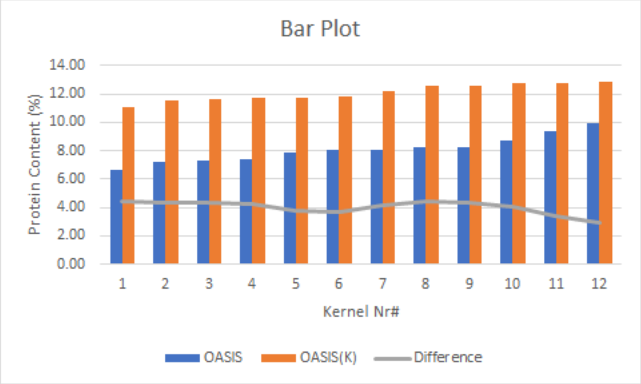
Figure 19: Tukey's multiple comparison test. X1: Parts of kernel, X2: sorts (0 = A75, 1 = OASIS)

Shown is the confidence interval for the means of protein content and significant difference between the two sorts seems to be a couple of percent points. Unfortunately, this is not enough evidence to showcase if this is systematic variation or not but is a good indication that FIA can measure the protein content difference on a significant scale. Overall, the mean protein values for OASIS and A75 are respectively, 8.10 % and 10.53 % whereas the values for found in Kjeldahl are 11 % and 15.45 %. One cannot confirm any systematic variation based on only these means.

To confirm a systematic trend, one must include the mean protein of all available sorts and compare to the method of Kjeldahl, even though it is quite time consuming.

However, additional results for Kjeldahl protein was provided for 12 OASIS kernels. One can therefore directly compare the difference between the 12 kernels from this thesis and the 12 kernels from the contacts of CSORT from Altai University. Sorted from lowest to highest the means are the following, *cf. Table 10*.

Table 10: Kernel protein of FIA, Kjeldahl and their difference. Bar plot added as illustration.

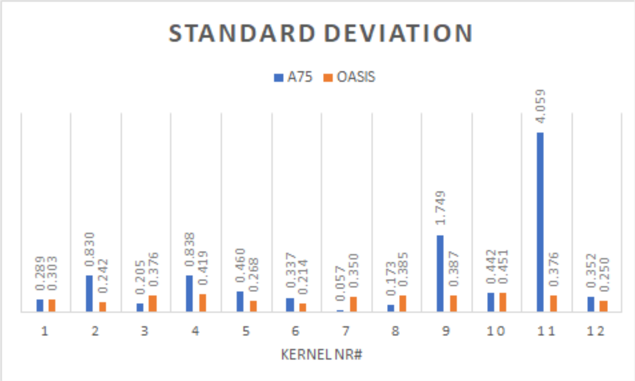
FIA	**Kjeldahl**	**Difference (mean: 4.004)**
6.67	11.06	4.39
7.19	11.51	4.33
7.32	11.63	4.31
7.44	11.7	4.26
7.91	11.71	3.80
8.06	11.77	3.72
8.07	12.23	4.16
8.20	12.59	4.39
8.24	12.59	4.35
8.67	12.72	4.05
9.35	12.74	3.39
9.93	12.84	2.91
		

The mean difference between protein for FIA and Protein for Kjeldahl seems to be stable around 4, this is somewhat of an indication that a systematic variation has occurred for this chosen wheat sort. Further experiments for other sorts are needed to confirm this maybe constant trend of a difference on 4. It is found later that the trend is indeed different and has a more linear fashion.

On a positive note, the method of FIA seems to give quite stable results when observing the standard deviation for each kernel with its three powders, *cf. Table 11*. If an outlier is present one can remove the subsample and still base the mean on two powders instead of three.

Table 11: Standard deviation for each kernel. Numbers marked in red indicates that an outlier is present in one of the powder samples.

Sort	1	2	3	4	5	6	7	8	9	10	11	12
OASIS	0.303	0.242	0.376	0.419	0.268	0.214	0.350	0.385	0.387	0.451	0.376	0.250
A75	0.289	0.830	0.205	0.838	0.460	0.337	0.057	0.173	1.749	0.442	4.059	0.352



Now that this method is deemed relatively stable and there is evidence that the difference is systematic, an even larger upscaling for all wheat sorts were conducted.

5.3.6 FIA: Sampling and Results of Preliminary Experiments - Part III

The main objective of this preliminary experiment is to confirm the observed systematic variation regarding protein of wheat kernels between the method of FIA and the method of Kjeldahl. This is done by including all given wheat sorts, calculate their protein mean content and compare them to the method of Kjeldahl to see if anything systematic is present. Since this is the only objective and precision is not the focus, the results from previous experiments, the 12 kernels form OASIS and A75 are used with the new results. Regarding sampling for other sorts, it will be smaller to save time since the preparation is quite time consuming when preparing three samples from one kernel. The complete sampling list of wheat kernels are shown below, *cf. Table 12*.

Table 12: Sampling list for preliminary experiments PART III.

SORT	OASIS	PAM	ALTSTEP	SALUT	ALTZH	ALT530	ALT105	ALT75	ALT325	SIBRIS
NR# Kernels	12	6	4	4	4	4	4	12	4	6

Now, since many different wheat sorts have been used, the kernels have been chosen and analysed in random order. This sampling gives a total of 60 wheat flours prepared and analysed.

The important statistics for the experiments on the kernel mean protein basis and a comparison with Kjeldahl are as follows, *cf. Table 13* and *cf. Figure 20*.

Table 13: Important statistics regarding protein content for FIA PART III with Kjeldahl protein (KP) for comparison. The table is sorted with respect to K.

SORT	Protein range	Mean	STD	Mean (K)
OASIS	6.79 - 9.93	8.09	0.90	11
PAM	7.20 - 10.59	8.88	1.44	12.21
Altstep	8.79 - 10.20	9.26	0.66	14.15
Salut	8.67 - 10.96	9.77	1.2	14.28
Altzh	7.54 - 11.25	9.45	1.52	14.4
Alt530	6.45 - 9.95	8.92	1.65	15
Alt105	9.14 - 9.89	9.61	0.33	15.2
Alt75	8.32 - 12.28	10.53	1.09	16.45
Alt325	8.82 - 12.10	10.19	1.38	16.55
Sibris	8.86 - 12.45	10.34	1.16	16.6

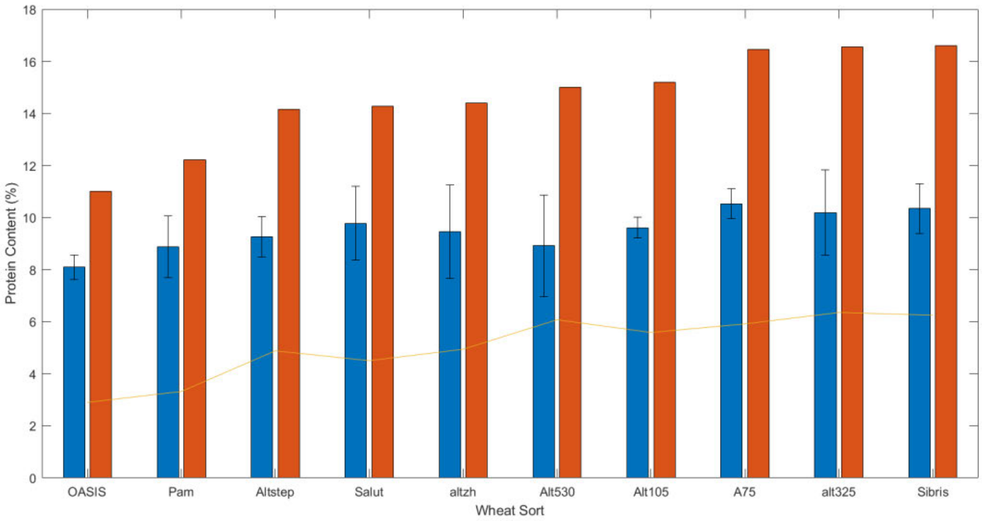


Figure 20: Bar plot of mean protein content of FIA (blue) vs Kjeldahl (red). Errorbars are 95% confidence intervals. The line resembles the values for difference between the two methods.

Three conclusions can be drawn from this plot:

1. The method of FIA can give very good confidence intervals when building the mean on 10+ kernels (see OASIS and A75)
2. All protein content measured with FIA are underestimated compared with Kjeldahl.
3. The difference between the two methods seems to increase in a linear fashion for the spanned protein range.

Making a linear regression between the predicted values of Kjeldahl with respect to FIA gives the following relationship, *cf. Figure 21*. With the model's ability to explain ~80% of the variation it is hereby confirmed that FIA underestimates protein content in a linear fashion, the variation is systematic.

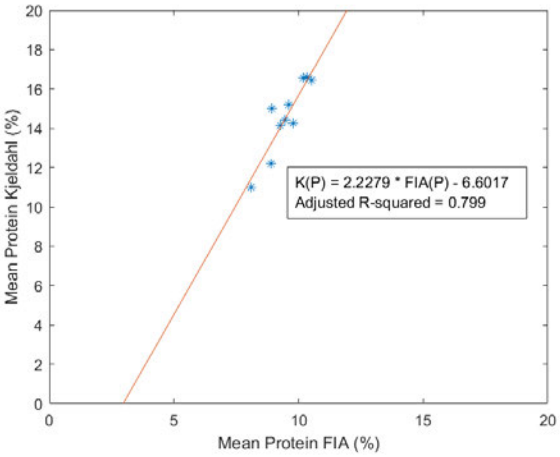


Figure 21: Linear regression model between the mean protein content of FIA and Kjeldahl.

Now it is interesting why there is a systematic underprediction. The problem may lay in the digestion procedure. The persulfate digestion is normally done on water samples with suspended solids. No studies could be found using this digestion procedure on wheat. It could be theorized that a fraction of the pulverized kernels would contain shell flacks that would be too big to be fully digested and therefore cause this underprediction. However, this would not explain the systematic increase in difference between the two methods.

Furthermore, a report [24] suggested some interferences without much detail. It mentions that there may be nitrogen compounds present that may be resistant to the persulfate digestion procedure. Also, it states that organic carbon can make a reaction with the persulfate reagent to generate CO_2. The report mentions that concentrations over 150 mg-C/L can deplete the digestion solution before all nitrogen have been oxidized and this would give a systematic underprediction [24]. This may be the case, but it's hard to evaluate without knowledge of the concentration of carbon.

To understand this variation fully, it may be necessary to conduct a full-factorial design on factors such as concentration of digestion solution, autoclave time, weight of kernel sample or degree of pulverization. Because of time limit, this was not possible.

Because of the *systematic* underprediction, very good stability and time savings compared to the method of Kjeldahl FIA was nevertheless chosen as the reference method to measure protein content in wheat kernels moving forward.

5.4 Hardness Determination

It was not possible to measure relative hardness of wheat kernels with the standard procedures because of two reasons 1) the equipment was not available and 2) the standard methods would make it impossible to measure vitreousness on the wheat kernels afterwards. To able to measure all parameters the kernels need to mainly be intact. Instead, it was thought out that using the rupture force of wheat kernels would be a sufficient representation of hardness. This could be achieved with a simple compression set-up. The inspiration to try out this method was found in the study [39]. When the break happens at the shell, it is theorized that a small amount of damage would happen to the endosperm of wheat making it still possible to observe the vitreousness of the wheat. When using this method, it is important to make a preliminary analysis of the influence of physical factors. Factors not related to the chemistry of wheat kernels could be the size of the wheat kernels or the orientation. The to-be sampled wheat would be categorised into three classes: Small (S), weight under 40 mg, medium (M) weight between 40 - 50 mg, Large (L) weight over 50 mg. They were all weighed.

Furthermore, two orientations were chosen the side orientation (O1) and the crease down orientation (O2), as seen below, *cf. Figure 22*. It is hypothesized that the crease down would have a higher rupture force because of its ability to stabilize itself. Additionally, one may wonder if the most accurate measurements would be in crease down orientation since this limit the kernels ability to move around.

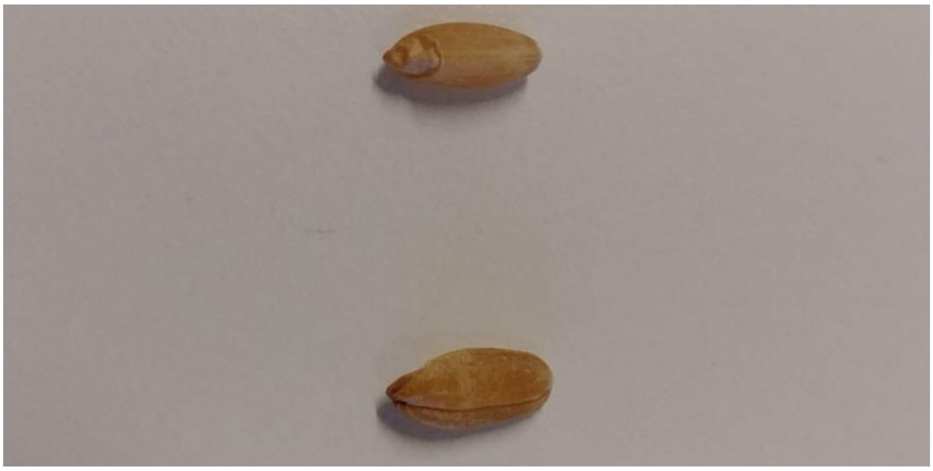


Figure 22: Representation of orientations. Top kernel is in crease down orientation (O2) and the other in side orientation (O1).

If the weight influences the rupture force it could be favourable to instead use the force/weight ratio for model building. If an orientation influences the rupture force all kernels would be carefully placed in a chosen orientation for the model-building experiments. It was later thought out to test dimension size with rupture force (broadness ("thickness") and longitude of the kernels), however no experiments were conducted on these factors.

All the preliminary experiments have been randomized and the compression speed was 2.00 mm/min throughout the experiments. The compression apparatus was automatically set to stop when encountering a force drop. However, this was not efficient, so most of the time the hardness measurement was manually stopped. Afterwards, each compressed kernel would be inspected to see if a crack in the shell had appeared, if not, a new measurement test on the same kernel was conducted. It was possible to crack the shell of most of the kernels without making visible damage to the endosperm.

5.4.1 Hardness Determination - Sampling and Results

40 kernels were chosen from the OASIS sort, 20 were to be crushed in O1 orientation and 20 in O2 orientation. The results can be showcased in the appendix. The following boxplot gives a summary of the results, *cf. Figure 23*. First discovery is that the range of the boxplots are very large spanning in around 100 N from the minimum to the maximum value.

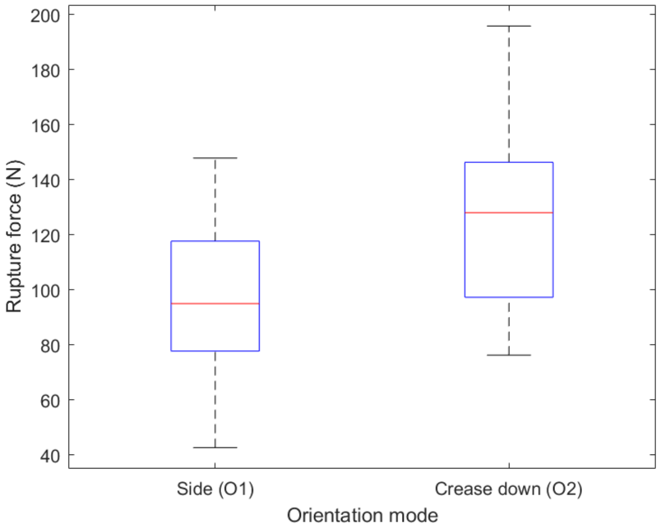


Figure 23: Boxplot of the 40 OASIS kernels, 20 in each orientation mode.

Second discovery, there seems to be high difference between the two means of the orientations. Using ANOVA one can now calculate if anything significant has occurred, *cf. Figure 24*.

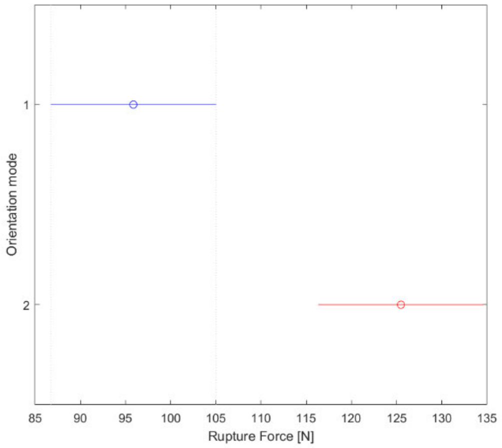


Figure 24: Multiple comparison analysis on the ANOVA results between the two orientations. Lines are 95% intervals for the true mean.

Indeed, the two orientation modes give significant different rupture force results, it is therefore very important to place the model building kernels in the same orientation. The side orientation (O1) seems favourable as it gives lower rupture forces and therefore minimizers the measurement time. There are also no obvious advantages in choosing one or the other orientation mode in regard to accuracy. One should be wary with this method as it's almost impossible to measure all the micro orientation states of the kernels, which possible gives the large variation in rupture force. Here the definition of micro orientation is a slightly deviation in the two chosen modes ex. a kernel in O1 orientation which is tilted because of its geometry would go under micro orientation mode. A bigger analysis would be needed but its outside the scope of this thesis. There is already a bit of scepticism for this method because it is then not possible to correct for the external parameters of micro orientation. The analysis and usage of this method continues for research purposes only.

Now, the effect of size (weight) of the kernels on the rupture force was tested. The sampling consists of 36 OASIS kernels all placed in the O1 orientation. As a quick reminder, the categories were as following: Small (S), weight under 40 mg, medium (M) weight between 40 - 50 mg, Large (L) weight over 50 mg. For the set of 36 kernels, 13 were small, 15 were medium and 7 were large (1 outlier). The summary of the force rupture results compared to the previous (20 kernels in O1) are as following, *cf. Figure 25*. This is done to underline there is a good repeatability for the distribution of datapoints.

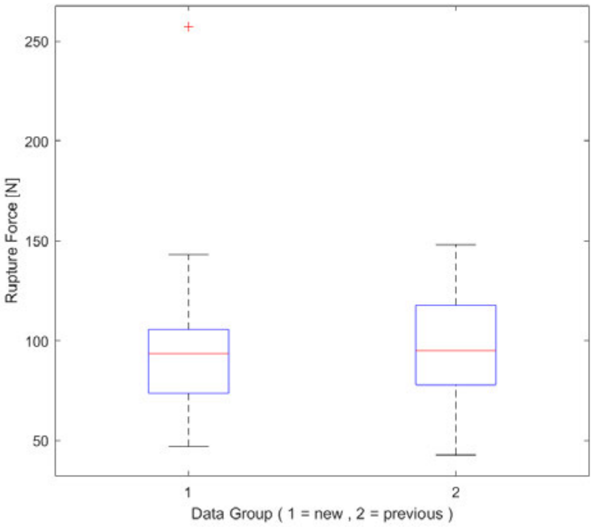


Figure 25: Boxplot of rupture force for two groups consisting of 36 kernels for group one and 20 kernels for group 2. The red mark is an outlier.

It is now very interesting to test, when collecting both datasets into one, to see if they are actually normal distributed for this given orientation. This is done with a QQ-plot and an Anderson-Darling test for a total of 55 kernels (outlier removed). The Anderson-Darling test returns a decision for a null hypothesis that the kernels data are normally distributed, the calculated p-value is 0.8408, the test then fails to reject the null hypothesis. In conclusion the data is normally distributed.

The QQ-plot is illustrated in *Figure 26*. For the most part it looks normally distributed, however, in the left tail there are some deviations form normality. There is no current explanation for this phenomenon.

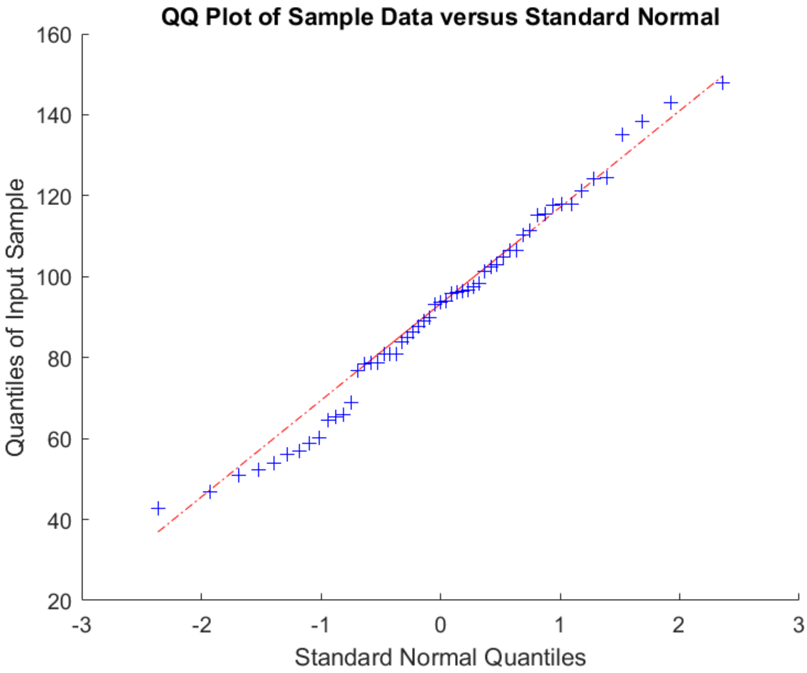


Figure 26: QQ-plot for the rupture force of the collected 56 kernels from OASIS in side orientation.

Regarding the weight, when plotting weight vs. rupture force the following summary can be shown *cf. Figure 27 and Figure 28*. For small and medium sized kernels, the rupture force seems random, however for kernels of larger size they tend to have higher rupture force, even significant under the current definitions. Overall the correlation coefficient can be calculated to 0.32 indicating only weak correlation.

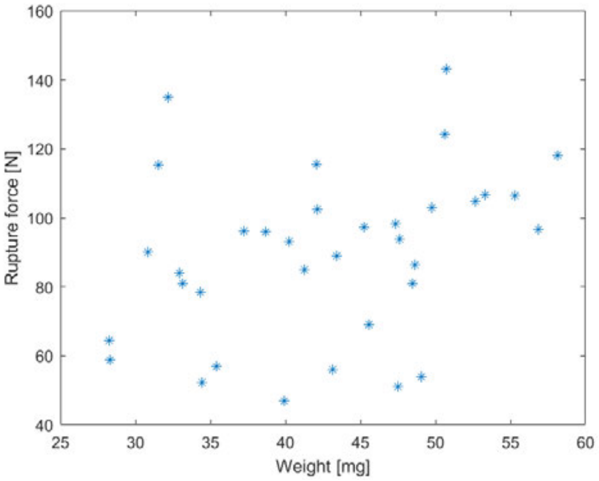


Figure 27: Rupture force vs. Kernel weight.

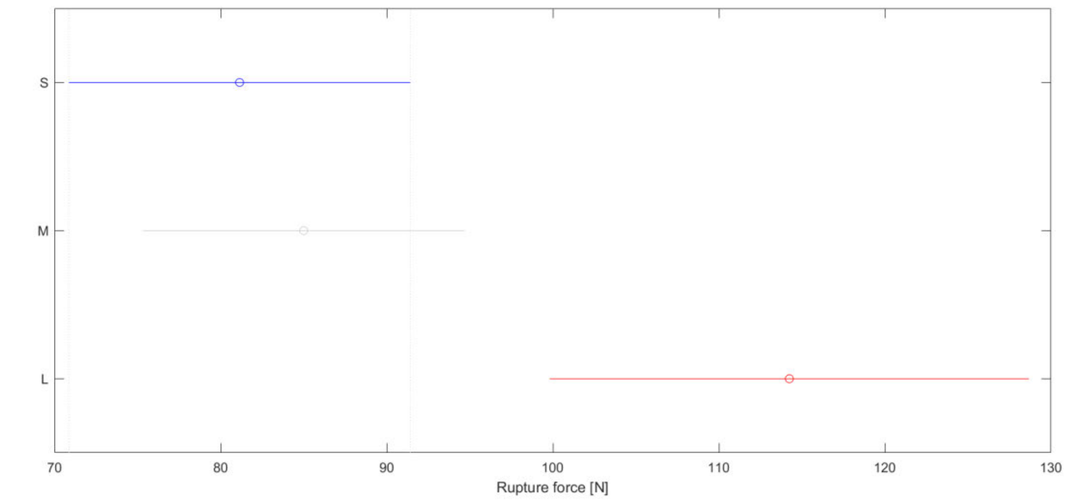


Figure 28: Summary of ANOVA results for rupture force and size classification.

To summarize, it has now been showcased that the rupture force is both dependent on orientation and size. It is thought that taking all kernels in one orientation and calculating the rupture force/weight ratio instead of rupture force in model building could lessen the problems associated with this reference method.

To see, how all kernel sorts compare, in this method, 12 kernels from each of the ten sorts were randomly chosen and compressed, all in "side orientation". The following boxplot displays the results, *cf. Figure 29.*

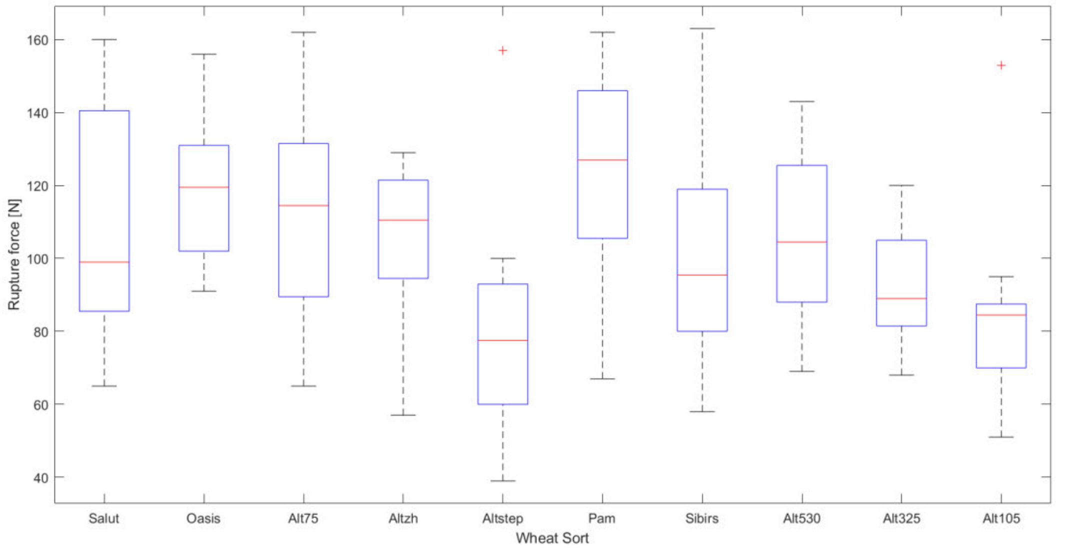


Figure 29: Boxplot of the 10 wheat sorts and their rupture force.

There are three worrying phenomena's in this boxplot. Firstly, for the most part, the ranges of rupture force seem to be very high. Secondly, there is evidence that there is correlation (r = 0.42) between wheat relative hardness and protein content [57], this phenomenon is not clearly seen here. Ex. OASIS and pam have the lowest means of protein (11 and 12 % respectively), however they seem to have the highest means of rupture force and Sibris and Alt325 (16.6 and 16.55 % protein content respectively) have the lowest mean rupture forces. But no significant change is seen between the mentioned sorts OASIS, pam, sibris and ALT325 *cf. Figure 30.* Thirdly, this method is more prone to outliers compared to the protein determination method of FIA, where the results were quite stable and no outliers were normally detected.

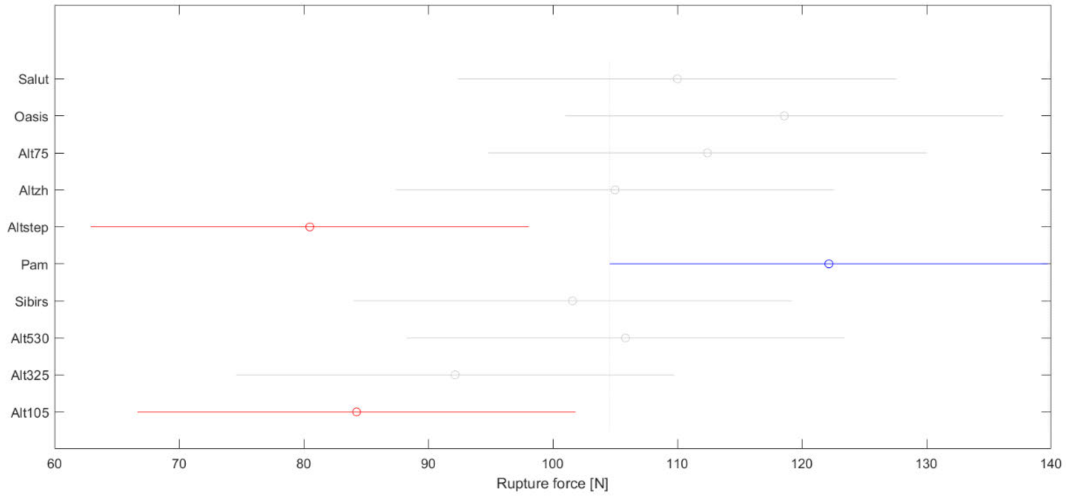


Figure 30: Comparison of rupture hardness for all wheat sorts.

The relative standard deviations for these measurements range from 16.55 - 37.29 % whereas relative hardness determination with NIR-spectroscopy can be as low as 9.9 % with regard to relative standard deviation [57]. There are therefore no high expectations with respect to accuracy of a model predicting rupture force.

However, since no other method was available to measure relative hardness of wheat without making vitreousness measurements impossible this method was chosen, with scepticism, as the reference method to measure relative hardness ("rupture force") and actions such as placing all kernels in one orientation and normalizing with weight could properly give okay predictions.

6 Main Experiments

Here the procedure and sampling will be presented for the main experiments.

As a reminder, the wheat for the main experiments consists of ten different wheat sorts. Their names, their type and their mean protein content can be summarized in *Table 14*.

Table 14: Wheat sort names, wheat type and Kjeldahl mean protein % (KP%)

Name	Alt75	Alt105	Alt320	Alt530	Altzh	Altstep	OASIS	Pam	Sibris	Salut
Type	Soft	Soft	Soft	Soft	Soft	Soft	Hard	Hard	Soft	hard
KP%	15.45	15.2	16.55	15	14.4	14.15	11	12.21	16.6	14.28

150 kernels were chosen to undergo image acquisition, hardness measurements, vitreousness measurements and protein determination. The methods were conducted in the order showcased.

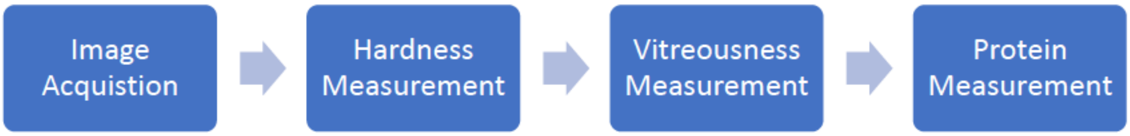


Figure 31: Schematic representation of order of measurements.

To represent each sort equally, 15 kernels from each sort were randomly chosen to give the total of 150 kernels. Furthermore, 100 kernels of the 150 were chosen for calibration and 50 kernels were selected as the test set for the models. The calibration and test kernels were selected with stratified sampling. In stratified sampling the whole sample size was divided into wheat subgroups. Now, for each subgroup (*cf. Table 14)* 10 kernels were randomly chosen for calibration set and the rest of that subgroup (5 kernels) was chosen for the test set. This sampling method ensures that each wheat still is represented equally in the calibration and test sets.

To ensure that the randomized order of the 150 kernels were the same throughout all the experiments they were placed in two sample boxes an 8 x 10 box and a 7 x 10 box. As an additional insurance random kernels from the two boxes were weight measured again between all major quality estimations to ensure the same order. It can hereby be concluded that the random order of the 150 kernels were the same throughout the experiments.

6.1 Image Acquisition

The model of the apparatus, taking the images, was a headwall photonics 1002a-00371, *cf. Figure 32*. The scanner used was a line-scan system with the subtype, whisk broom scanner. The scan length was set to 9 cm with a 40 µm distance between the pixel lines. The spectral resolution is unfortunately only in the NIR interval of ~938–1623 nm, whereas the complete NIR-region is 780–2500. Between 938–1623, every fifth wavelength was measured giving 142 wavelengths (variables for the chemometrics part) in total. The spatial resolution of the line camera was (2251x320). In other words, 2251 lines of a single image was taken with each line consisting of 320 pixels. Overall, this setup would generate hypercubes of the size: (2251x320x142). However, the hyper cubes are smaller under data processing because of removal of background and bad pixels.

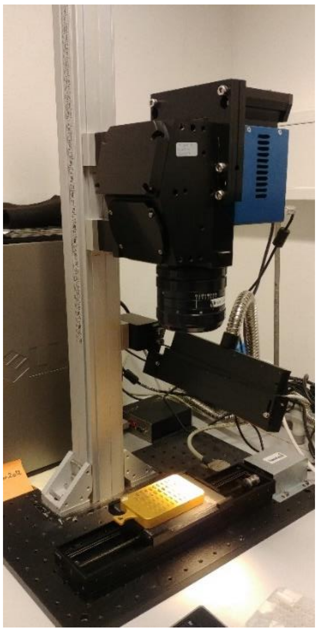


Figure 32: Scanning set-up.

Images for four orientations of the 150 kernels were taken. 16 hypercubes (images) were made in total for the 150 kernels. The orientations include crease down, crease up and the two sides, *cf. Figure 33*. Each kernel was manually rotated 90° (in the same order as in *Figure 33*) and carefully aligned doing the image acquisition to acquire the images for the four orientations.

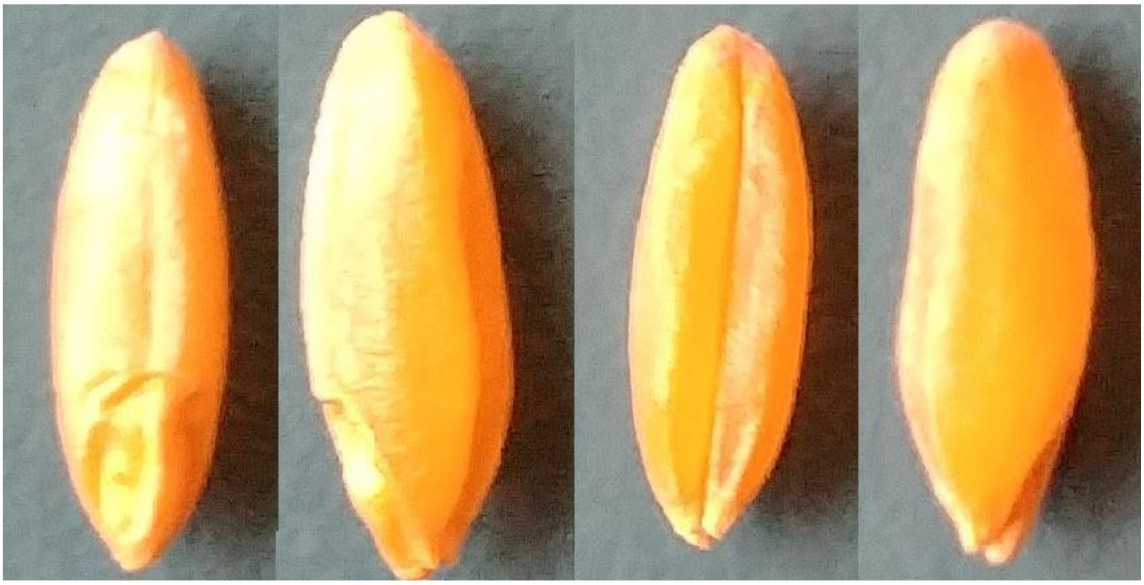


Figure 33: The four orientations of a random OASIS kernel.
From left to right: Crease down, first side, crease up and the second side.

Before image acquisition, the apparatus was calibrated with a white background and then a black background. For the white calibration normal white paper was used and for the dark calibration, a black cloth was upheld to block out the light. The software used the references in a standard way, such that each new image is subtracted from the dark reference image and thereafter normalized with the white reference image with respect to the following formulae:

$$X_{corrected} = \frac{X_{measured} - X_{darkref}}{X_{lightref} - X_{darkref}}$$

The background used for kernel imaging was a metal holder. Each holder could contain up to 50 kernels, *cf. Figure 34*.

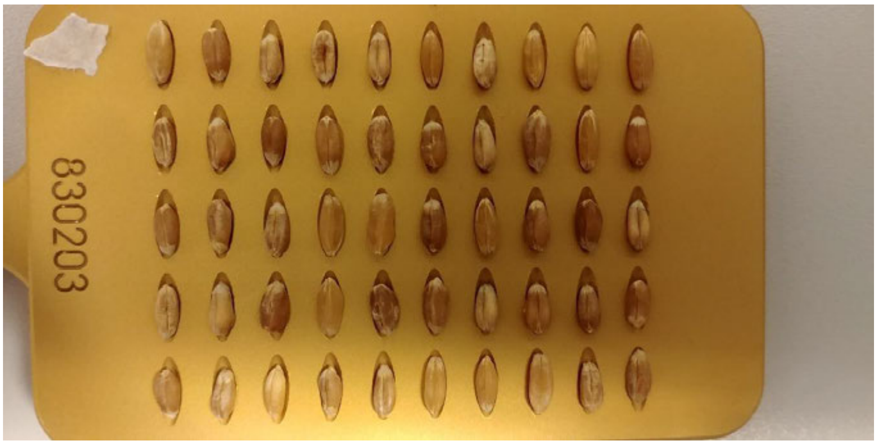


Figure 34: Kernel metal holder with kernels in crease up orientation.

It was investigated that a typical kernel has an average pixel amount of 1871 pixels without erosion based on a 30-kernel image, therefore, the cumulative feature is based on average 1871 pixels for each kernel and lower depending on the erosion magnitude.

The background pixels for each image were removed by a deterministic rule. Firstly, each image was reshaped to a [number of pixels X wavelength] matrix. Afterwards, the pre-processing of SNV was conducted on this matrix and the method of PCA (principal component analysis) was made to determine the characteristics of each pixel. To review the characteristics, the scores plots were observed, however only for the first PC since it explained close to 80% of the variation. An example of a score plot for a 50-kernel image is illustrated below, *cf. Figure 35.*

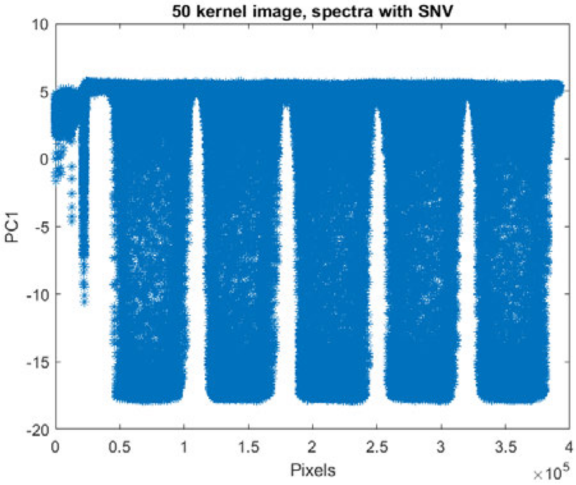


Figure 35: The pixels score plot for PC1.

Now, from *Figure 35* a threshold was chosen of zero, such that all pixels under this line were excluded. This was set as the masking operation for each image. One could also choose a slightly different threshold, but when reviewing the mask, this threshold seems fine, *cf. Figure 36*.

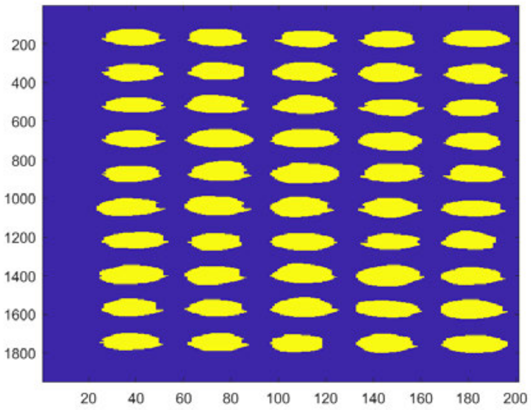


Figure 36: Mask with initial threshold.

The data-files from the apparatus containing the hypercubes was converted to data-files easy to access with Matlab with a custom-made script made by professor Jose Manuel Amigo Rubio.

6.2 Data Review and Initial Preparing

All data, except the hypercubes, used for model generation are in attachment BIG DATA. If there is special interest for the hypercube files, please refer to the preface. Every kernel has gotten a unique ID associated with it and every orientation has gotten an ID, 1 is the top orientation, 2 the first side orientation, 3 the crease up orientation and 4 the second orientation. In the dataset there will also be presented each kernels sort, their weight [mg], their protein content [%], their relative hardness ("rupture force") [N], their vitreousness type and an experimental value called vitreousness value [%] which is how much percent of the surface is non-vitreous. Initial models have been built on these data to know where to set the settings. Settings are here defined as thresholds for VIP/SR and number of latent comps. It will be explained in more detail moving forward. Outliers were removed from the protein content data in the Excel script otherwise possible outliers were removed in the Matlab scripts for both vitreousness and rupture force.

The protein content of the kernels was measured batch wise. As a reminder, the kernels were placed in two boxes (8 x 10 and 7 x 10). From the preliminary experiments and from a practical viewpoint the maximum number of samples to run in a single FIA batch were 60 samples. In the chosen method each kernel was split into three parts, so 20 kernels (2-rows) could be measured in one batch. Such a batch took between two - three days to prepare and measure making it possible to only run two batches in a week. The procedure was the same as in the preliminary experiments. For each batch a 1.66 mg/l glycine, a blank and a 5 mg/l NO_3 were added to ensure stability. Furthermore, for each batch, the FIA were calibrated and recalibrated for each ten sample with a 5 mg/l NO_3.

It was observed at start-up that the calibration would not measure a 5 mg/l NO_3 correctly properly because of temperature variations inside the apparatus. To combat this problem the FIA apparatus was calibrated two times at start-up with a 15-minute interval between each calibration. The second calibration was stable. The stability checks can be reviewed in appendix of further interest. However, the means for the blank was 0.033 mg/L NO_3-N, for the glycine validations 1.824 mg/L NO_3-N and for the 5 mg/L test 4.948 mg/L NO_3-N.

The measured blanks showcase a very little present of nitrogen contents which mostly can be because of uncertainty of the calibration at such low nitrogen content levels. The glycine solutions however showcase a slightly higher concentration than the 1.66 mg/l. This variation can be because that normal measuring cylinders were used to measure water, digestion solution and glycine solution from the stock solutions. This was done in order to trade off accuracy for higher speed when making the samples of the wheat and glycine. The difference was not foreseen to be this large if this is indeed the case. It may therefore have been better to at least use precision equipment for the glycine solutions. However, it is strange that there is a systematic increase for all glycine solutions from the 1.66 mg/l and at the time of experimentation this was evidence that the digestion was completed fully. It is acknowledged that the glycine solutions shouldn't have been measured with normal measuring cylinders and with the current procedure the glycine solutions may not even be valid. The 5 mg/l solutions had nitrogen concentrations slightly lower than 5 mg/l but the difference was seen to be at acceptable levels.

For the spectrometric data the cumulative feature to be used as response values are the mean spectrums for all pixels for a unique kernel. Here it is also possible to combine images. A combination of an image includes calculation of each images mean spectrums for the kernels separately. Then, for each unique kernel, the two mean spectra for them are averaged to give a new mean-of-means spectrum as a cumulative feature for a combination image. 11 images, single or combinations, are chosen to give the features. The combinations can be reviewed below, *cf. Table 15*.

Table 15: Image combination table.

Single images

- Crease down
- Crease up
- Side 1
- Side 2

Double image

- Crease down and side 1
- Crease down and crease up
- Side 1 and crease up
- Side 1 and side 2
- Crease up and side 2

Every-orientation

- Side 1, crease down, side 2 and crease up

Several statistics are important to consider when evaluating which model is the best. The model should not be too complicated therefore it is preferable to have the least amount of principle components for the model and secondly, the R-squared for cross validation should be as high as possible to give a good indication of model quality.

Initially for each model generated these statistics were collected:

- Number of latent variables.
- Coefficient of determination, R-squared.
- Coefficient of determination for cross-validation, R-squared CV.
- Root-mean-square error, RMSE.
- Root-mean-square error for cross-validation, RMSE CV

As a reminder, data for 150 kernels are present, however, the models are only built on 100 kernels and 50 kernels are used as a test set for the generated models. The maximum number of latent components were set to 15.

6.3 Full-Factorial Design for Experiments

Each mathematical model for prediction can be made on several factors. It is therefore necessary to make a full-factorial design to take all these factors into account and see, which have the biggest effect on the prediction power. In other words, one most search for the best model. A sub-method of full-factorial design is used called grid search. Grid-search identifies the most important factors by using a grid with fixed combinations and tests every possible combination. The factors to be included in the grid search are:

1. The degree of erosion (4).
   Here four levels are chosen: no erosion, 1-pixel, 2-pixel and 3-pixel erosion. The structuring element for the erosion was chosen to be disk-shaped and here the definition of the levels corresponds *to the radius of the disk-shaped structuring element.* For reference the 1-pixel and 2-pixel (erosion structuring element) are showcased here *cf. Table 16*.

   Table 16: Magnitude of the erosion levels with regard to structuring element.

Erosion level	Structuring element
1-pixel	$\begin{matrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{matrix}$
2-pixel	$\begin{matrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{matrix}$

   It is hard to evaluate how much of the kernels contour consists of bad pixels, by using erosion on different levels one could find the most optimal level of erosion to remove these pixels.

2. Image combination (11).
   There are currently 11 possible image combinations, *cf. Table 15*. Image combination is used as a factor because of the possibility that using the mean-of-means spectrum for multiple images one could stabilize the spectrums giving better predictions.

3. Pre-processing (5)

   Five different common pre-processing techniques have been used to eliminate the additive effects (baseline shift) or multiplicative effects (slope of spectral data). They can be summarized in the following table, *cf. Table 17*.

Table 17: Pre-processing levels for grid search.

Pre-processing Technique	**Effect on response values (X)**
No pre-processing (1)	No effect
SNV (2)	Centring and scaling of spectra to correct the baseline.
MSC (3)	Adjust all spectra to a mean spectrum correcting both the slope and baseline offset.
Savitzky-Golay (SG) polynomial derivative filters (1st derivative) (4)	Removal of noise and correction of spectral slope.
SNV and SG (1st derivative) (5)	Removal of noise and correction of spectral slope and baseline.

On a side note, a 2nd derivative SG pre-processing technique was carried out on some initial trail models for protein content prediction. This technique didn't seem to improve the prediction ability and the method is therefore excluded. A combination of SNV and SG is instead carried out.

4. Variable selection (3).

   Three levels of variable selection were chosen to adjust the number of variables and only take the most important into account. The levels are 1) no variable selection 2) VIP threshold selection 3) SR threshold selection.

   For each of the three quality parameters a custom set of general VIP and SR thresholds were manually designed. This was done by building an initial trial model for prediction on each quality parameter. Afterwards the models were manually improved by looking at plots of VIP and SR followed by an adjustment of the VIP/SR thresholds to improve R-squared for cross validation. Another trail model was developed to test if the threshold also was okay for others.

   Now, there is a problem with this procedure, that is, that the chosen thresholds are only very precise for the generated initial models. This proved to be a problem in the model generation when a model didn't have many important variables (bad predicting model). This caused errors in the generation script because of exclusion of all variables by the threshold. To avoid excluding to many variables from other models, the thresholds were chosen somewhat conservatively.

   When the best model is found a manual improvement on the VIP and SR threshold will instead be made.

The following figure, *cf. Figure 37* , summarizes the grid search.

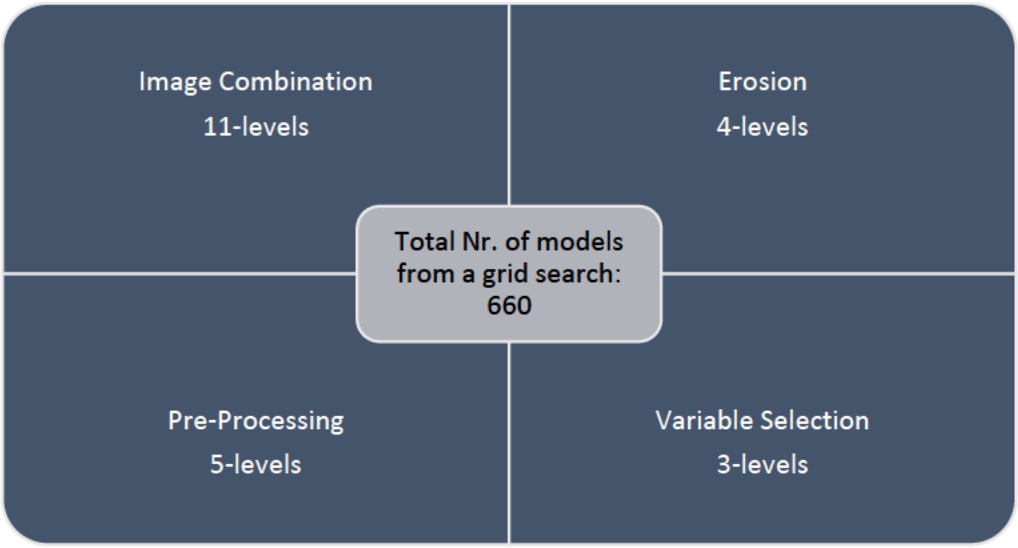


Figure 37: Summary of levels in the grid search.

Multiple manually made custom Matlab scripts are generated to make the following grid search, a new one for each quality parameter. Included in the scripts are a manually made add-on that can switch between finding the statistics for min R-squared for cross validation and finding Wold's R for a threshold of 1.

7 Presentation of Models: Review and Discussion

Since the method of PLS makes latent components based on both the X and Y-space a single model can be made to predict all three quality parameters simultaneous. However, the chosen approach is to individually build models on every single quality parameter. This gives the following advantage that its clear to see which reference method is giving the best prediction. If for instance a reference method builds bad models it would be timewise wasteful to try to build models combining this bad prediction with ex. a good prediction. It can already now be unveiled that vitreousness and rupture hardness are badly predicted therefore no models predicting multiple variables are showcased.

Making vitreousness prediction models based on cross-sectional area [%] gives very bad predictions. Instead the vitreousness models are built upon classification where the two types to be classified are defined as non-vitreous and fully vitreous. Kernels showing only little vitreous behaviour are classified as non-vitreous. The classification is done by defining a binary variable for the vitreous data (response data) in which each class are either defined as +1 or -1. The PLS models are built upon this new dataset and predicted responses closer to -1 are predicted as one class and predictions closer to +1 are predicted to the other class.

By analysing the mean statistics for each quality parameters grid search one can now get an overview on what model's quality parameter on average gives the better prediction.

The following showcase are based on criteria one, minimum RMSE for cross-validation, *cf. Table 18*.

Table 18: Summary of model statistics from the three grid searches. Nr. of latent components are rounded up.

	Mean: Nr. Latent components	**Mean: R-squared**	**Mean: R-squared for cross validation**	**Mean: RMSE**	**Mean: RMSE for cross validation**
Protein Prediction [%]	8	0.7962	0.7085	0.647	0.779
Rupture force Prediction [N]	6	0.3219	0.1648	29.82	34.00
Rupture force prediction weight ratio [N/mg]	6	0.2904	0.0890	0.6778	0.8061
Vitreousness Prediction (CLASS)	3	0.1925	0.0821	0.701	0.769
Statistics for models with max R-squared for cross validation					
Protein Prediction [%]	8	0.8427	0.7842	0.5708	0.6703
Rupture force Prediction [N]	14	0.6823	0.4237	20.54	28,82
Rupture force prediction weight ratio [N/mg]	15	0.8467	0.3433	0.3180	0.7137
Vitreousness Prediction (CLASS)	14	0.8405	0.4364	0.3153	0.6397

Looking at the averages of the grid search, it is quite clear that models that predict protein are quite good with the R-squared for cross validation being 0.7085 on average. The other parameters get predicted very poorly by their models. This were to be expected since the reference methods were subjective and sub-optimal compared to the standard methods. It is interesting that there appear to be "okay" models when looking at the models giving the max R-squared for these two parameters, hardness and virtuousness. It is now interesting to see, which factor and level of the factor, had the best performance on prediction ability in general and what combination the best performing models have.

7.1 Main Effects and Combinations of the Grid Search

Before evaluating the best combinations, there will be an evaluation of the two criteria's for choosing the optimal number of components, criteria 1 with minimum RMSE for cross validation and criteria 2 for Wold's R. The evaluation will be based on protein prediction since it gave the most stable models. A summary of the levels for the factors are included in *cf. Table 19.*

Table 19: Summary of factors for grid search. SE = structuring element.

Levels	**Erosion (radius of SE)**	**Picture combination**	**Pre-processing**	**Variable selection**
1	No erosion	Crease down	No pre-processing	No variable selection
2	1. pixel	Side 1	SNV	VIP
3	2. pixel	Crease up	MSC	SR
4	3. pixel	Side 2	Savgol	-
5	-	Crease down side 1	SNV + Savgol	-
6	-	Crease down Crease up	-	-
7	-	Crease down Side 2	-	-
8	-	Side 1 Crease up	-	-
9	-	Side 2 Side 4	-	-
10	-	Crease up Side 2	-	-
11	-	Every	-	-

For criteria 1 (minimum for RMSE CV) the main effect plot is shown in *Figure 38*.

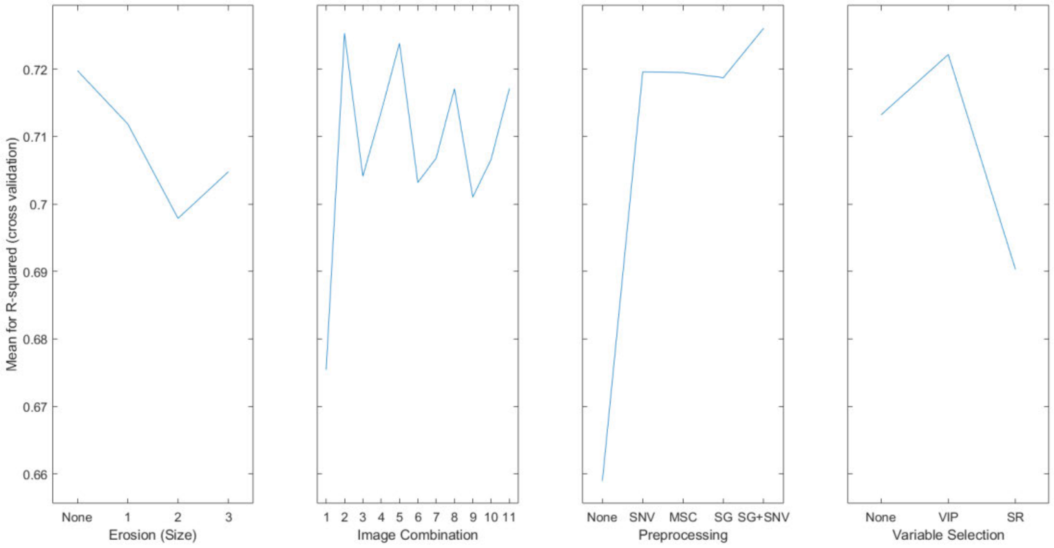


Figure 38: Main effect plot for criteria 1 based on mean of R-squared CV.

For criteria 2 (Wold's R threshold) the main effect plot is shown in *Figure 39.*

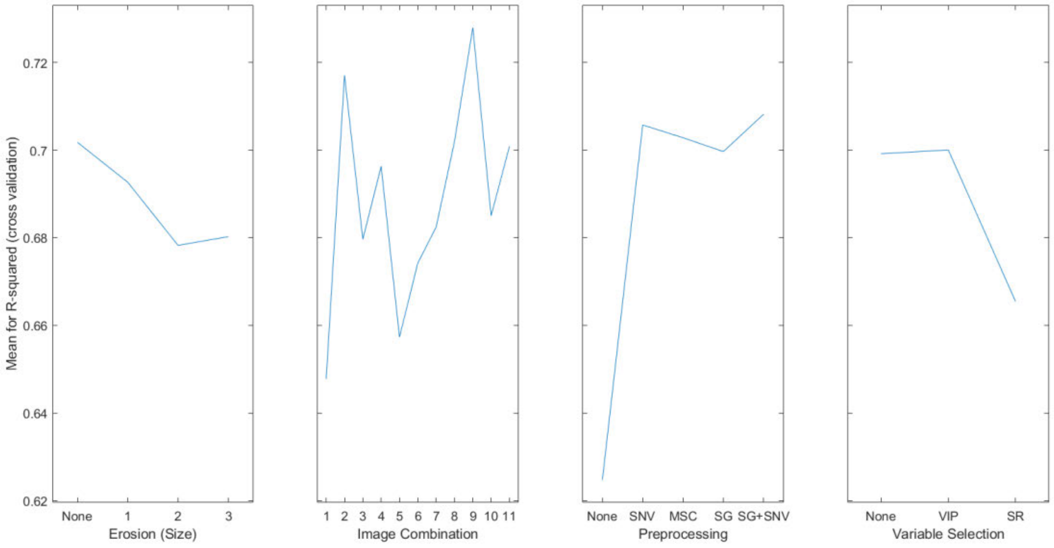


Figure 39: Main effect plot for criteria 2 based on mean of R-squared CV.

It is clear, that there are lot of similarities between the two criteria to find the optimal number of components based when looking at R-squared for CV. Criteria 1 seem to give slightly higher R2 values, a one-way ANOVA with null-hypothesis that the means are the same confirms with a p-value close to zero that there indeed is significant difference between them rejecting the null-hypothesis. Criteria 1 significantly gives slightly higher R-square values. A multiple comparison can be seen in the appendix.

Now returning to the main effects, *Figure 38 and Figure 39*. It seems that more erosion gives worse prediction models possibly because of the removal of good pixels for both criteria. A strange phenomenon happens that a slight increase in prediction ability happens between erosion level 3 and 4, could it be that there are isles of bad pixels in the kernel images that gets removed with level 4 erosion?

The biggest difference between the two main effect plots are the image combinations. Criteria 2 seems more chaotic while criteria 1 seems to have better models when the models are built on images of the kernel sides. Regarding pre-processing a similarity is seen between the two criteria's, the Sav-Gol deviate filter with SNV seems very promising for this type of data. Regarding variable selection a likewise similarity is seen. The third level of variable selection seems to not improve the models. However, this is because the threshold for level 3 (selectivity ratio) was set very conservatively.

Another point to see in the grid search criteria is how many principle components each criterion gives. It was significantly found that criteria 1 gives around ~1-2 more principal components then criteria 2. The multiple comparison plot between the two criteria can be reviewed in the appendix.

It's hard to confirm what criteria is superior based on this evidence but if more complexity doesn't matter criteria 1 seems better. Also, not showcased here, criteria 2 are very troublesome to implant in the grid search if the models have very bad prediction ability. The threshold would almost always terminate after a couple of latent components while criteria 1 seems to complete the whole model generation cycle with more latent components. So, for convenience and since no confirmative conclusion can be made on which is better, the Hardness and vitreousness models will be made with Criteria 1.

The combinations for the best models are as following built on criteria 1 for the three factors.

Table 20: Best models (Max R-squared CV) and their combinations.

Quality parameter	**Erosion**	**Picture combination**	**Pre-processing**	**Variable selection**
Protein	3-pixel erosion	Combination between two sides	Sav-gol	VIP
Hardness	No pixel erosion	Side (2) single	Sav-gol	VIP
Vitreousness	2-pixel erosion	Combination of crease up and crease down orientation	Sav-gol + SNV	VIP

These models will now also be manually improved, and a more aggressive SR/VIP threshold will be tried out and finally the test set will be included to the final models.

7.2 Best Models for Protein Prediction

Now, building the same model using the combination in *Table 20* different VIP/SR thresholds were tried out to see if the model could be improved. Furthermore, different filter sizes on Sal-Gov were tested to see if the model could be improved. Changing the filter size didn't improve the model, the optimal thresholds were found as the following, *cf. Figure 40*.

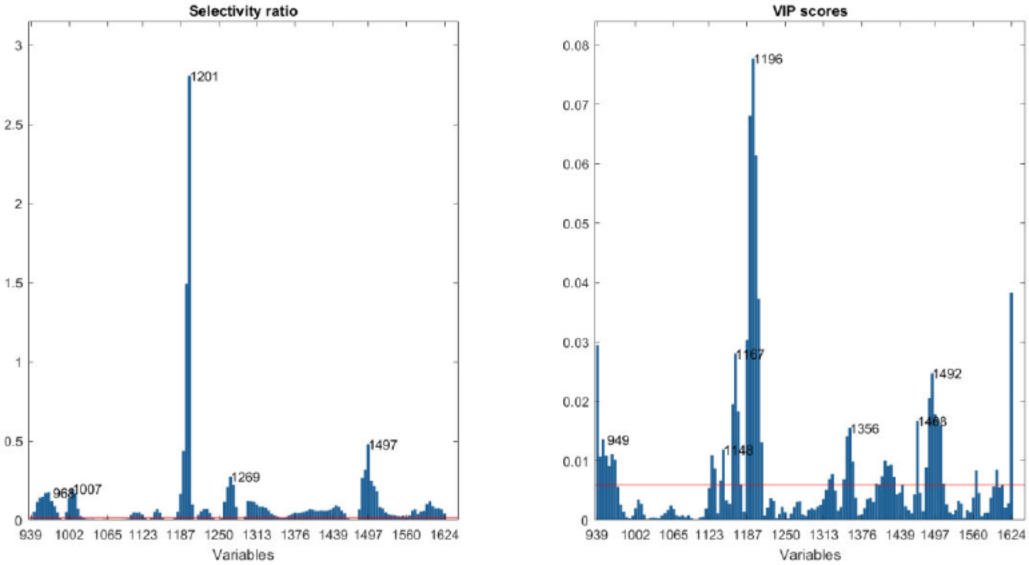


Figure 40: Selectivity ratio and VIP scores. Red line is the optimal threshold found. Biggest peaks are included. Be aware of different Y-axis!

It is interesting to give a chemical evaluation of the most important peaks to see which functional groups they correspond to. A summary of the chemical interpretation are as follows, cf. *Table 21*.

Table 21: Chemical interpretation of the most important wavelengths.

~Wavelength [nm]	Chemical interpretation
950	CH third overtone and OH second overtone. (H_2O)
1000	NH second overtone
1200	CH second overtone.
1370-1440	OH first overtone, NH first overtone, CH combination first overtone $CCNH_2$/CONHR groups CH_3, CH_2 groups.
1500	RNH_2 (NH first overtone)
1600	ArCH

The most important wavelengths resemble a specimen that contain carbon, nitrogen and oxygen holding groups with carbon being the most predominant this is most likely protein seen here.

After small adjustments to the VIP-threshold a summary of the model with the test set can be presented with 8 principal components, *cf. Figure 41* .

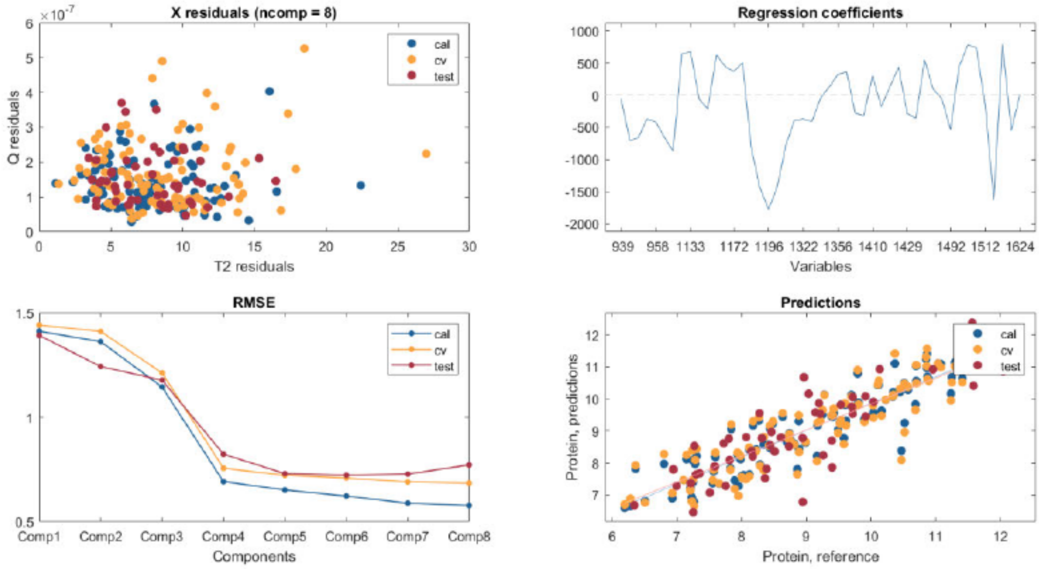


Figure 41: Summary of the best model.

Looking at the residuals, it is evaluated that no outliers are present. Furthermore, the predictions of the calibration, the cross-validation and test-set are quite good. The summary of raw statistics is likewise presented in *Table 22*.

Table 22: Core statistics for the best protein predicting model.

	R-squared	**RMSE**
Calibration	0.839	0.578
Cross validation	0.775	0.685
Test set	0.712	0.772

But how does this model compare to other studies? When using only Near-infrared spectroscopy, studies have found a very high level of prediction for the bulk of samples with R-squared for validation of 0.95 [58] which is quite impressive. However, using only a near-infrared camera the disadvantage is that the internal variation of wheat kernels cannot be analysed, and it is impossible to analyse wheat kernels individually. The research done on the topic on using HSI in protein prediction of wheat are somewhat limited, but two studies, [14] and [16], have tried a similar approach. The main difference is the reference method used to analyse protein. The study, [14], used the Dumas combustion method for determination of protein content, while [16] used a semi-micro Kjeldahl method. Furthermore, there are also changes in wavelength intervals and sample size, which will be reviewed shortly.

A table can be presented displaying the main statistics of the studies and this master thesis, *Table 23*.

Table 23: Summary of model statistics.

	Nr. Comp	R2	R2CV	RMSE	RMSECV	Interval [nm]	Cal. size	Test Size
Master Thesis	8	0.839	0.775	0.578	0.685	938–1623	100	50
[14]	13	0.824	0.773	0.868	0.958	980–2500	1647	(522)
[14]	14	0.732	0.681	1.063	1.129	1000–1700	3138	(992)
[16]	10	0.855	0.852	0.721	0.539	928–1695	57	22

The generated best model for protein prediction in this master thesis compares very good with the two other studies. Using Kjeldahl as reference method seems to produce the best prediction model but using FIA as the reference method seems to produce better models than dumas combination (even in a smaller wavelength interval). It can hereby be concluded that FIA can stand in as an attractive method for measuring protein and using it as a reference method with HIS when predicting protein in wheat. One disadvantage is that a systematic underprediction is present however, as found in the preliminary experiments. There is also evidence from [14] that using a bigger spectral resolution can improve the prediction models but come at a greater economic cost. In [14] the improvement in R-squared for cross-validation is ~0.09, so by using a bigger spectral resolution the protein prediction with FIA could theoretically be improved to 0.86 for R-squared CV.

Under image acquisition, the orientations of the kernels in the two studies are based on crease down orientation in [14] and random orientation in [16]. There is evidence in this master thesis, that measuring the kernels in side orientation seems to be favourable. This may also be a reason that the R-squared values compare favourably with the other models even in a smaller wavelength interval and even with less latent components. On a side note, the prediction quality may not be sufficient in commercial use as the project partner, CSORT, preferred better prediction models.

Furthermore, one can also showcase the predictions as a concentration map and hereby illustrate the protein content for individual kernels, *cf. Figure 42*.

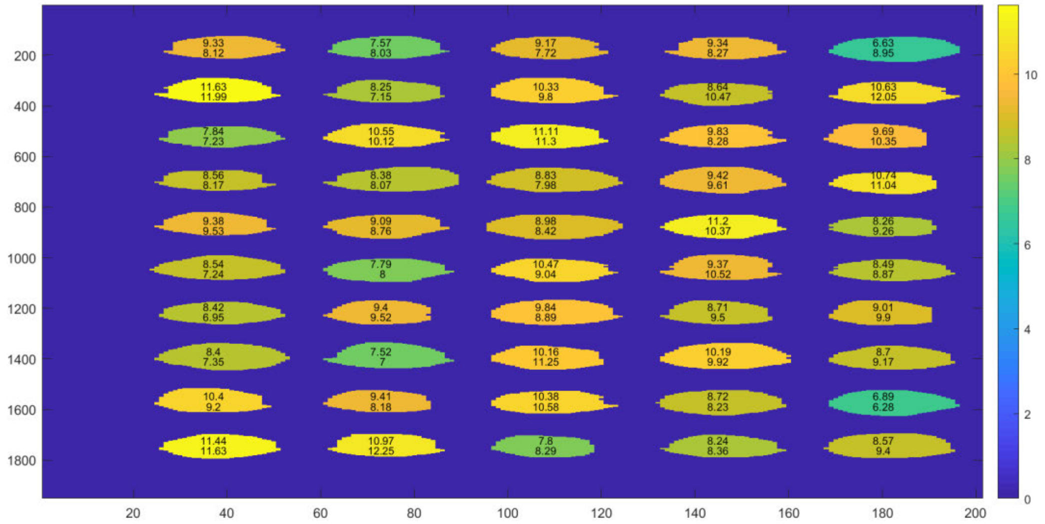


Figure 42: Application of the PLS regression model on an image with 50 kernels, orientation is side1 no erosion. Colour bar is made on protein content (%) for the predicted values (top values). FIA reference values are located as the bottom value for each kernel.

When predicting the same kernels in crease down orientation a slight difference in protein is seen on an individual level, however, the concentration map appears more or less the same, *cf. Figure 43*.

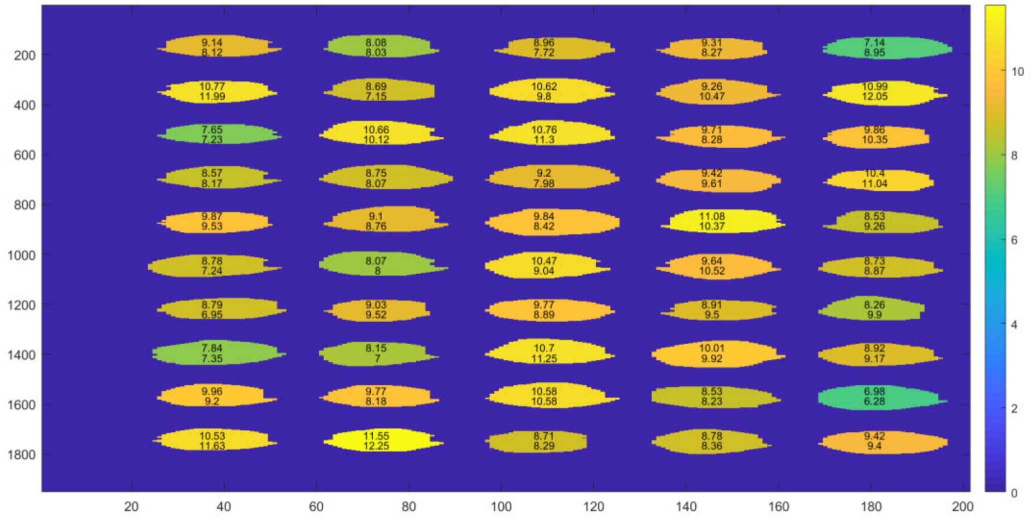


Figure 43: Application of the PLS regression model on an image with 50 kernels, orientation is crease down no erosion. Colour bar is made on protein content (%) for the predicted values (top values). FIA reference values are located as the bottom value for each kernel.

This means that the model in general gives the same prediction indications independent on kernel orientation.

7.2.1 Protein Prediction Using Random Forest

Another method for generating a regression model other than using PLS is to use the random forest procedure. A random forest model consists of an ensemble of decision trees. Decision trees are predictive models that based on observations for an object's variables can go down different nodes (feature decisions) to conclude and give an estimate on the real value (in their leaf). A single tree tends to overfit their training set, however using a forest, "a bag" of decision trees, reduces overfitting [59]. In this procedure a sub-method is used to create Bootstrap-aggregated (bagged) decision trees. Here multiple training sets are generated from the original training set, they are made uniformly and with replacement from the original set. An equal number of models (trees) are created based on the number of training sets. The trees are collected and the output of all their final decisions is averaged to give an estimate of, in this case, the protein content [60].

To ensure no overfitting and more generalization, 3000 trees have been grown. Firstly, the variables (mean NIR spectrums for every kernel) have been pre-processed with Savitzky-Golay transformation (1st derivative) and variable selected with a VIP threshold, the image is a combination of the two sides with 3-pixel erosion. This combination was first chosen since it gave the best prediction model with the PLS method. The training set used for "calibration" of these trees are still the 100 kernels used when making the PLSR. One way of evaluating the prediction ability of a random forest is to evaluate the "out-of-bag error", for regression trees this corresponds to mean squared error for them. The following figure shows the out-of-bag error development for the 3000 trees grown, *cf. Figure 44*.

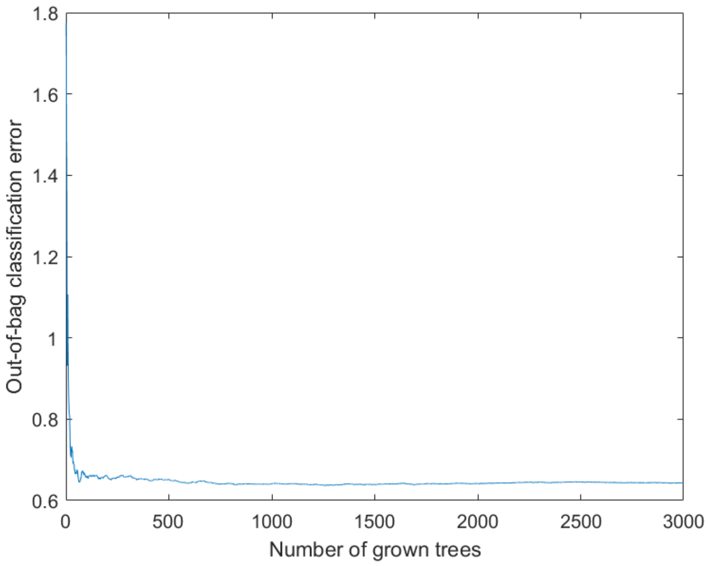


Figure 44: Out-of-bag error development for 3000 trees. For 3000 trees, the out-of-bag classification error translates to R^2~0.7.

It is clear, that after 1000 trees the model doesn't improve much.

Before showcasing the raw statistics, it is worth mentioning that two other variable selection methods arise with random forest. One can look at what variables is selected the most for each decision, summed up for all trees. The more a predictor is selected the more important it is. The second criterium is permutated importance estimates of each variable. This is calculated for each tree by altering “permuting” each variable decision threshold, if the variable has high influence altering its value will cause high error. A difference between the normal error and this error are calculated and a mean is found for the predictor variable, finally by division with the standard deviation it is possible to get this importance value. For the best model in predicting protein the importance variables and chosen threshold are, *cf. Figure 45*.

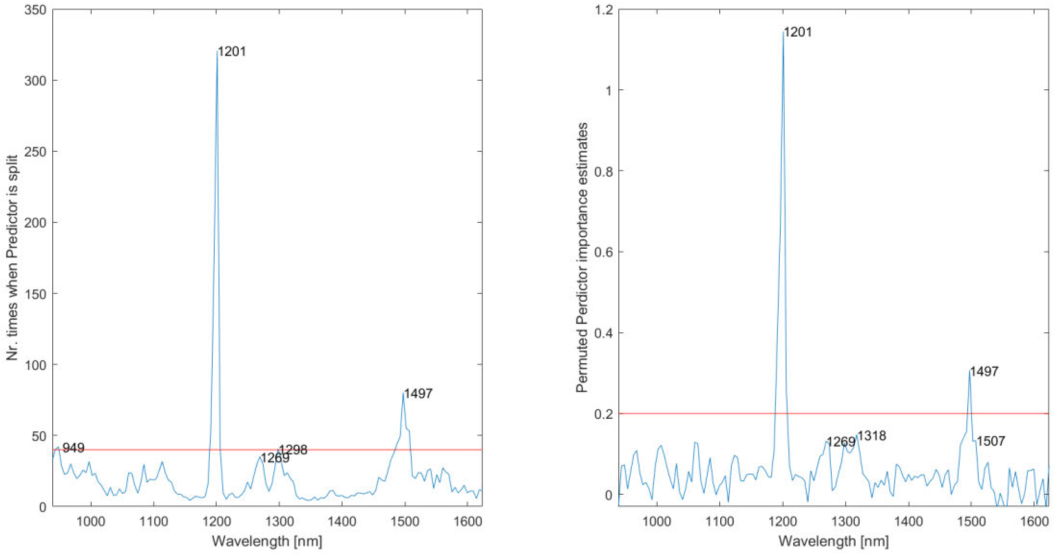


Figure 45: Overview of most important variable (wavelength) for the random forest model. Left plot the basis is number of times split, the right plot the importance estimates.

In comparison with the VIP and SR threshold from the PLS method two peaks are very similar the 1201 peak and the 1497 peak otherwise the smaller peaks are different for VIP and SR thresholds see *Figure 40*. It also seems that random forest behaves most optimal with a lesser number of variables compared to VIP and SR where more variables are included.

The thresholds have been found by testing an interval (for nr. split) between 10 - 60 with 10 steps. The best generated models seem to be produced at the threshold 40. For the permuted estimates the threshold was tried out varying from -0.05 - 0.25 with the best models being generated at a threshold of 0.20.

Different initial models have been made on different criteria to explore the statistics of the random forest. These can be reviewed in *Table 24*. The statistics for the 50-kernel test-set are also present.

Table 24: Summary of statistics for random forest procedure for specific conditions. The best model is highlighted.

Image combination & orientation	**Pre-processing**	**Variable selection**	**R^2**	**RMSE**	**R_{TEST}^2**	**$RMSE_{TEST}$**
side1 & side2	Sav-Gol (1st)	VIP	0.708	0.778	0.736	0.721
Top	Sav-Gol (1st)	VIP	0.594	0.917	0.733	0.733
side1 & side2	SNV	VIP	0.505	1.01	0.442	1.05
side1 & side2	Sav-Gol (1st) + SNV	VIP	0.683	0.811	0.71	0.754
side1 & side2	Sav-Gol (1st)	PIE	0.716	0.767	0.741	0.713
side1 & side2	**Sav-Gol (1st)**	**NR. split**	**0.722**	**0.758**	**0.748**	**0.704**

The best model generated from this test set gives a R^2-validation of 0.748 around the same as the method of PLS. The random forest method could be an attractive alternative to build regression models predicting protein content in wheat, but further testing is needed to significantly confirm this.

On a final note, it is noted that the random forest method is treated unfairly in this thesis. If it should be treated fair a full factorial design with grid search should be made taking all factors and levels into account. Unfortunately, because of limitations of time and that random forest isn't the main method, this have been disregarded.

7.3 Best Models for Hardness (“Rupture Force”) Prediction

Presented below is the statistics and summary for the best rupture force model with an updated VIP threshold of 0.0090 for 14 latent components, *cf. Table 25*.

Table 25: Summary of main statistics for best rupture hardness model.

	R-squared	RMSE
Calibration	0.683	20.5
Cross validation	0.424	28.8
Test set	0.004	50.0

From this one can say that the generated model is badly overfitted. It can be concluded that rupture force cannot be effectively predicted with HSI. Most likely, it is the reference method with the external factors that causes this. The main effect plot from the grid search have been made but is not included in the master thesis since it isn’t meaningful to continue analysing this bad prediction.

A summary to showcase outliers and predicted vs. reference are showcased with three principal components. Three is chosen because it gave the best prediction of the test-set (~R^2 = 0.02) .

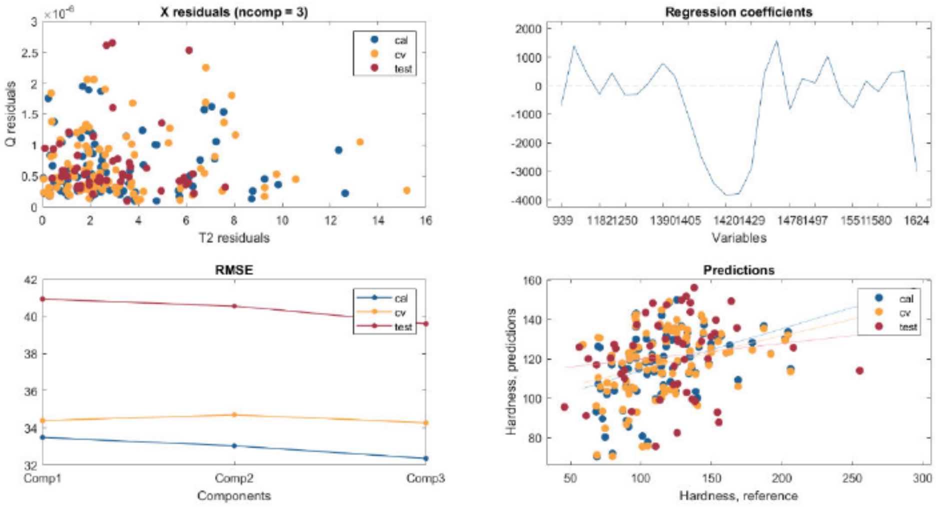


Figure 46: Summary of best model for rupture force prediction.

As seen, this prediction is mostly random. Rupture force with simple compression cannot be measured with HSI in this thesis. When calculating the correlation coefficient between a dataset of mean protein and rupture force no correlation is seen (r = -0.07), whereas in [39] a correlation of r = 0.953 was seen. Now it is interesting as to why this big difference have occurred even though the same method has been used. One big difference is the loading speed. In this master thesis the loading speed was 2 mm/s whereas in the study, the loading speed was only a fraction of this, 0.027 mm/s. The high speed could have made it harder to detect the actual rupture force. Moisture content also influence the rupture force, in the study [39] moisture content was accounted for and in this thesis moisture content was not accounted for. The moisture content could therefore prove to be valuable in hardness prediction with rupture force. This parameter could have been measured simply be oven drying but wasn't thought on during the experiments. A combination of the mentioned factors and a strong influence by orientation and weight could have caused these bad predictions viewed here.

Literature suggests that it indeed is possible to predict relative hardness of wheat kernels with HSI. A study [15] did achieve a correlation coefficient on R_{CV} = 0.8505 and $RMSE_{CV}$ on 0.1448 for a spectral range of 900 - 1700 based on PLS with reference method in hardness index of wheat. Therefore, it can be concluded that the chosen reference method in this master thesis not was optimal.

7.4 Best Models for Vitreousness Classification

As a summary the vitreousness (VIR) models were made using PLS discriminant analysis. All kernels were divided into two classes either fully vitreous or none-vitreous kernels. If a kernel showed small vitreous behaviour it was classified as non-vitreous. The best model was found using an image combination of the crease up and crease down orientation with 2-pixel erosion using both a VIP threshold and sal-gov (1st) with SNV. There were 18 kernels from the calibration set that has been excluded since it was not possible to measure vitreousness and 13 kernels from the test set that have been excluded. Most of the kernels were vitreousness in nature. The VIP scores and the chosen threshold for the most optimal model are shown in *Figure 47*.

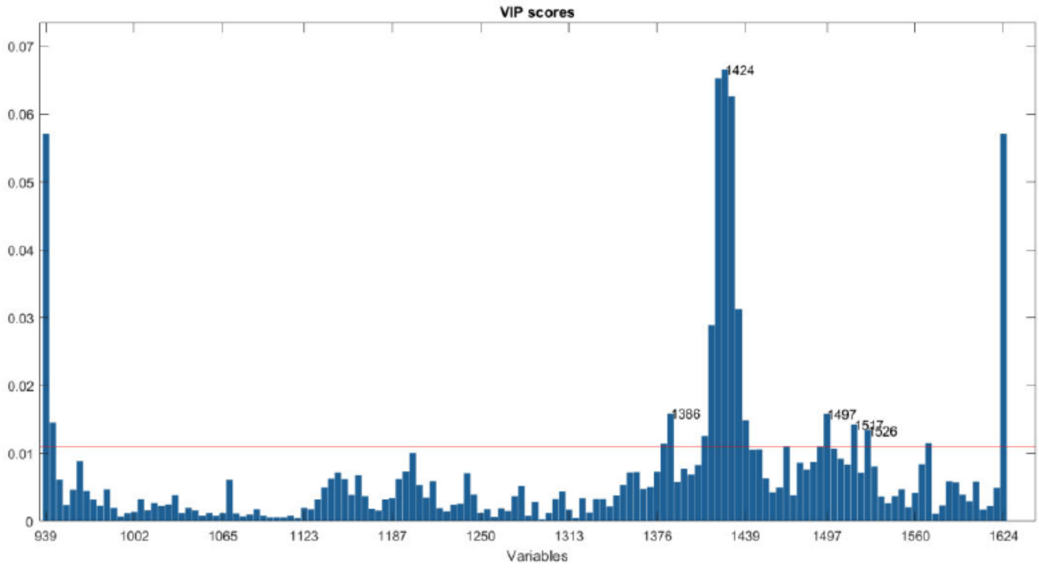


Figure 47: VIP-scores and threshold for the best VIR classification model.

Interesting the most important variables are around 1424 nm which is different from the protein prediction where the most important variables were around 1201 nm. Chemically around 1424 nm there are many overlaps between nitrogen, oxygen and carbon signals, therefore its hard to say anything on the chemical composition. Summary of main statistics can likewise be presented for calibration, cross-validation and the test-set for the best model with 4 latent variables, *cf. Table 26,* and the classifications can also be showcased, *Figure 48*.

Table 26: Statistics summary of best VIR-model.

	False negative	False positive	Sensitivity	specificity	Misclassification [%]	R^2
Calibration	3	1	0.75	0.9	4.88	0.51
Cross Validation	3	1	0.75	0.9	4.88	0.39
Test set	6	0	0.45	1	16.2	0.57

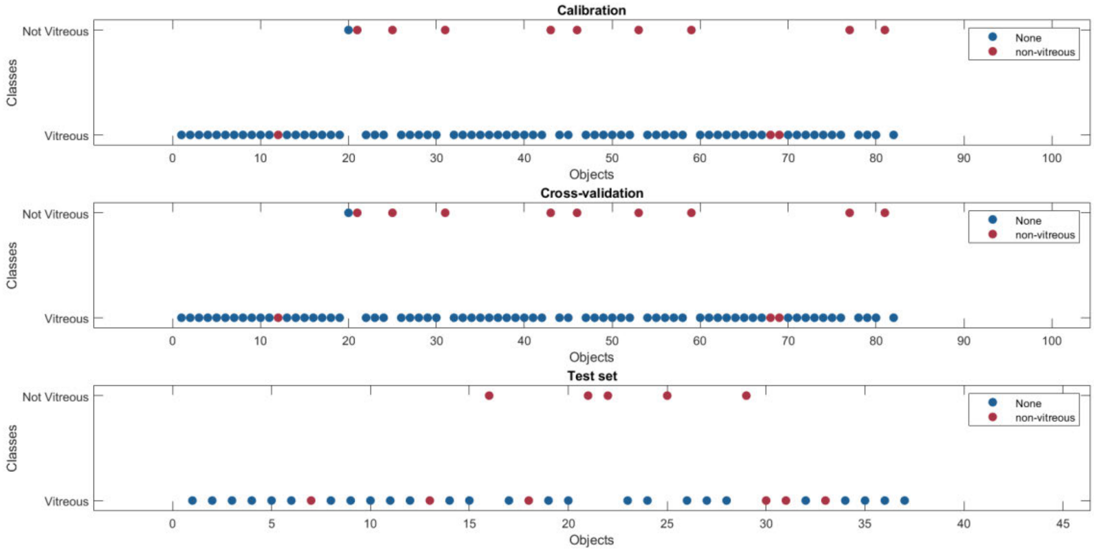


Figure 48: Showcase of classifications for calibration, cross-validation and the test set.

As one can see, many of the kernels were vitreous in nature. The results showcase that it is possible to develop a VIR-prediction model. The drawbacks of the experiments may decrease the prediction ability so that only an "okay"-model is made here. One of the drawbacks are the fractures in the kernels after hardness measurements. Sometimes the fracture lines would give the kernels a none-vitreous behaviour, when this was the case the kernels would be classified still as vitreous as a guess, this could explain the false negatives. Furthermore, a dataset that contain more not vitreous kernels would be preferable to see if the same tendency occur.

Another study, [17], using HSI for classification of vitreousness in wheat obtained perfect classification with two vitreousness classes (vitreous and non-vitreous) with the PLS-DA method and even obtained 94 % classification between non-vitreous kernel types in a wavelength interval of 650–1000 nm. Their sample consists of 600 kernels and vitreousness was measured using visual inspection and the special farinator cutter. Definitely, it is possible to predict vitreousness from the evidence of this study and based on the findings here one can conclude that there is also tendency that HSI can classify vitreousness type. However, it is recommended that vitreousness is measured with standard equipment without being damaged first with another reference method and that the operator should have more experience measuring vitreous beforehand with a bigger sample size of different vitreous classes to avoid bad prediction.

One could measure vitreousness and protein content instead of all three parameters for a potential better prediction model on vitreousness and protein. It is concluded that the experimental set-up for measuring vitreousness has not been optimal.

7.5 Prediction Based on Wheat Class

Together with Kjeldahl protein it was possible to get the classes of the ten wheat sorts. As a reminder, the data were as follows, *cf. Table 27*.

Table 27: Wheat sort names, wheat type and Kjeldahl mean protein % (KP%)

Name	**Alt75**	**Alt105**	**Alt320**	**Alt530**	**Altzh**	**Altstep**	**OASIS**	**Pam**	**Sibris**	**Salut**
Type	Soft	Soft	Soft	Soft	Soft	Soft	Hard	Hard	Soft	hard
KP%	15.45	15.2	16.55	15	14.4	14.15	11	12.21	16.6	14.28

Now it could be interesting to see if a PLS discriminant analysis with the HSI/NIR data could provide clear classification regarding wheat type. Each kernel of the 150-kernel dataset got classified according to the table in a binary format -1 = soft and 1 = hard.

Now using the same conditions as for the best protein predicting model, image side combination with 3-pixel erosion and sav-gol pre-processing but without VIP variable selection the following summary can be made, *cf. Figure 49*. The model is only built on the same 100-kernel model building set.

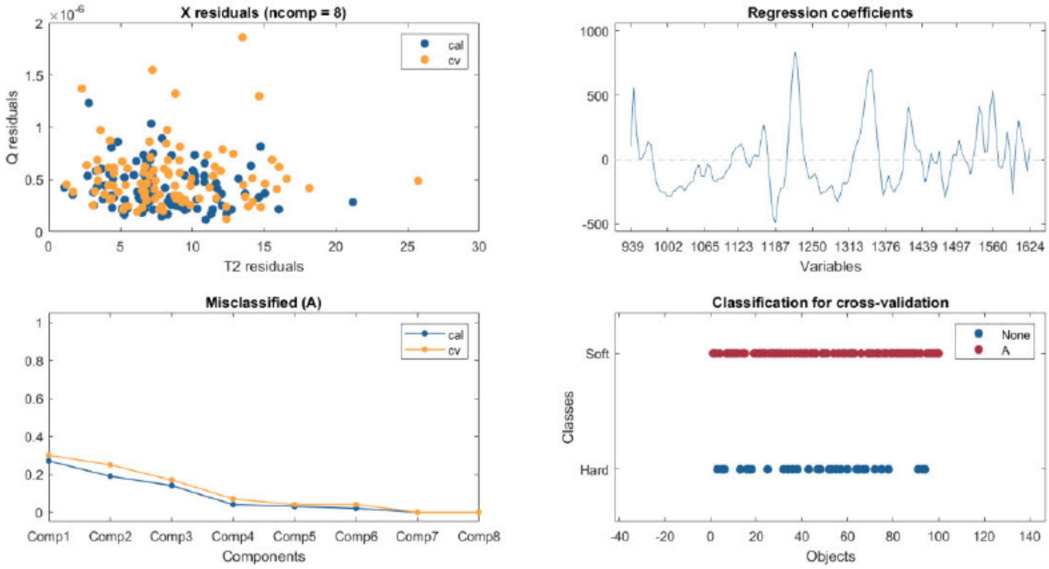


Figure 49: Summary of wheat type classification model.

As one can see, perfect classification between the wheat types occur for cross-validation for 8 latent components indicating that indeed wheat can be classified according to hardness class.

Including the 50-kernel test set the following summary can be seen, *cf. Figure 50*.

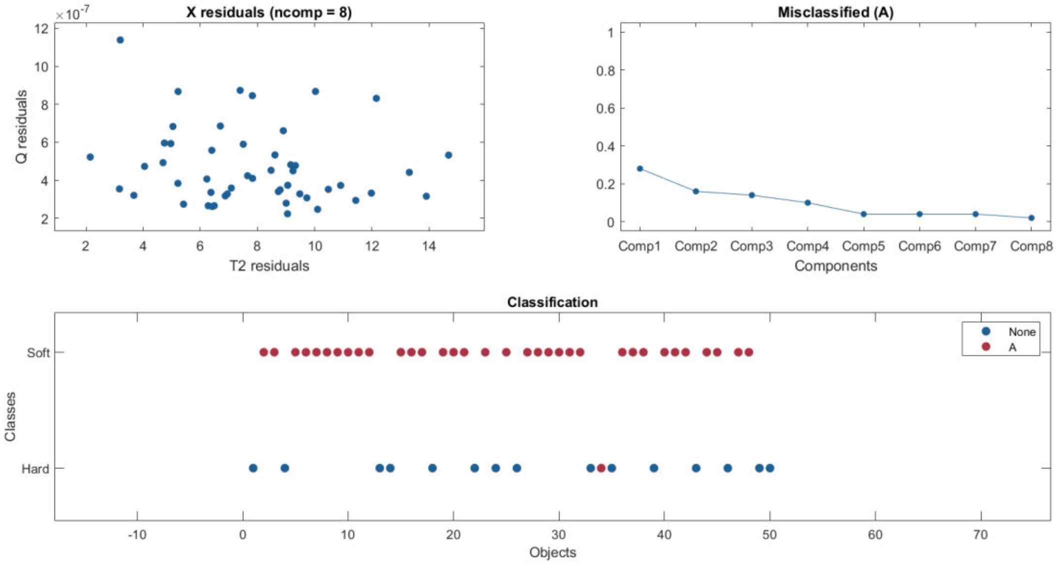


Figure 50: Summary of test set prediction for classification model.

More specifically, the statistics are the following, *cf. Table 28*.

Table 28: Summary of statistics for classification model.

Nr. Latent components	Explained X variance	Explained Y variance	False negatives	False positives	Sens.	Spec.	Mis
1	94.4	17.2	5	9	0.857	0.769	0.28
2	2.81	26.6	1	7	0.971	0.829	0.16
3	1.67	1.62	2	5	0.943	0.868	0.14
4	0.4	14.1	2	3	0.943	0.917	0.1
5	0.259	13.4	2	0	0.943	1	0.04
6	0.0847	9.29	2	0	0.943	1	0.04
7	0.0803	1.78	2	0	0.943	1	0.04
8	0.0848	1.67	1	0	0.971	1	0.02

So, the generated model has 98 % correct classification for the test set giving very good evidence that HSI with NIRS can give near perfect classification on hardness class of wheat kernels.

8 Conclusion

The main goal of this 1-year master thesis was to evaluate HSI in the NIR and its prediction on quality parameters of wheat. To achieve this goal several objectives were established. First objective was to evaluate what quality parameters to predict on the same individual kernels and what reference methods to use for reliable, time saving and precise prediction of the chosen quality parameters. The three quality parameters chosen were protein content, rupture force as a representation for hardness and vitreousness. These were chosen because of their importance, available literature and the possibility to measure all three parameters on the same kernels with proper reference methods. The best reference method for protein prediction was evaluated to be flow injection analysis (FIA) mainly chosen for time-saving reasons and lack of other equipment (DUMAS). Preliminary experiments found that FIA was very stable when measuring protein content on wheat kernels based on the mean of 3 flour samples from the kernel. Compared to the standard method of Kjeldahl a systematic underprediction was seen and it was not possible for time limitation to test why this happened. Rupture force was thought to stand in for relative hardness of wheat kernels and was measured with simple compression. The reference method was chosen as this was the only semi non-destructive way to measure relative hardness of wheat kernels, so it was thought. Preliminary experiments showcased that this reference method was plagued with external factors such as kernel weight and kernel orientation on a significant level. Because of research purposes this reference method was still used for the main experiments. Regarding vitreousness a standard method was used, visual evaluation classing, with non-specialized equipment.

The next objective was image acquisition and calibration of models to see which criteria produced the best models. The main experiments consisted of 150 kernels, all randomized between 10 sorts equally. 100 kernels were to be used for model building while 50 kernels were used as the test set. Hyperspectral images were taken of all kernels in four different orientations, two sides and top/bottom orientations. The NIR was ~938–1600 with 142 wavelengths the spatial resolution 2251x320. The cumulative feature chosen for the wheat kernels was mean NIR spectrum. Background was removed with a custom mask based on PCA. The mathematical method used was projection to latent structures for calibration of the reference data with the spectral data.

To evaluate different conditions a full-factorial design was made with 4 factors, erosion (4 levels), image combination (11 levels), pre-processing techniques (5 levels) and variable selection (3 levels) for each of the three quality parameters. The optimal number of components were found using two techniques minimum of RMSE and Wold's R. The minimum of RMSE was deemed sufficient as the difference between the two techniques were very small, however minimum RMSE gave slightly better models on a significant level. A meaningful analysis for optimal number of components could only be done in protein predicting models as they gave the best predictions. It was seen that more erosion gave worse models, therefore the prediction quality is dependent on the number of pixels. Better models were also made with the side orientations. Regarding pre-processing and variable selection better models were produced using standard normal variate and Savitzky-Golay Polynomial derivative filters and Variable Importance in Projection as variable selection.

The third objective was to judge if any of the reference methods or the HSI method was useful for making quality prediction models. The conclusion is as follows:

The protein prediction resulted in very good models with R^2 for cross-validation as high as 0.775. Using the random forest procedure instead of PLS showcased likewise results. In comparison with literature the results were very similar, and it was estimated that even better models could be produced with more expensive equipment. Flow injection analysis could indeed stand in as an alternative method for measuring protein in wheat kernels on an individual level.

The rupture force prediction models were not able to predict hardness of the wheat kernels even though literature has evidence that hardness of wheat can be predicted. The reason is clearly in the chosen reference method and experimental setup as to many external factors were present such as weight or shape.

The vitreous prediction models based on classification of two types of vitreous behaviour was “okay to bad” with 16.2% misclassification for the test set. The reason for bad prediction may be because the kernels were damaged after rupture force predictions and that suboptimal equipment was used making vitreous behaviour harder to evaluate. Literature could showcase a near perfect classification regarding vitreousness.

When classifying wheat kernels based on hardness class a perfect classification model was made.

In this thesis the main vision was to measure multiple quality parameters on the same kernels. However, it was clearly a bad idea to include rupture force as a hardness prediction. It is advised for further projects to only use two quality parameters for the same kernels.

Because of bad prediction a model predicting multiple variable was not produced.

In summary:

It is possible to characterize wheat kernels based on protein content and wheat hardness class. Following literature, it may also be possible to characterize based on relative hardness and vitreousness. However, inferior reference methods made it impossible to characterize wheat kernels based on relative hardness and vitreousness in this master thesis.

9 Further Work

In this master thesis the systematic variation of the reference method of for protein determination was accepted, however, it could be interesting to test why this systematic variation occurs in the first place. The main hypothesis is that the digestion of the wheat doesn't release all the nitrogen in the solution. This could be tested more thoroughly by making a factorial design and factors could include different concentrations of active solutions, alternative digestion procedures or other autoclave settings (more digestion time, 60 minutes instead of 30 minutes). The advantages would be that FIA could completely stand in as a semi-automatic method for determination of protein in wheat. In this master thesis, there is convincing evidence that this would indeed be the case.

One of the main flaws of the experimental set-up was that sub-optimal methods were used under testing of the three quality parameters. Rupture force could in this instance influence the observation of vitreousness. This gave very bad prediction of Hardness ("rupture force") and vitreousness. It is suggested that instead of using three quality parameters on the same kernels one could use two. By using two either protein content and hardness or protein content and vitreousness one could use established standard methods that would give much better prediction.

Additionally, it is thought that using a bigger wavelength interval spanning the whole NIR would give better predictions since more overtones and combination bands for nitrogen groups are seen. Better spatial resolution would also be interesting to test. However, this suggestion is discouraged without a strong economic foundation.

The focus when generating models for protein prediction has been to use the cumulative feature of mean spectrums for every kernel. However, as proposed in [61] it is possible to have another cumulative feature. The idea behind this feature is to use principal component analysis (PCA) to make a collected component space for the pixels of every object on an image. By doing so, these "pixel" spaces will in theory have unique patterns for different objects. Afterwards, one can do a quantitative evaluation on these patterns and make this the new feature for, in this case, protein prediction. The method is deemed useful when analysing difficult cases where a lot of pixels have similar properties as in wheat kernels [61]. This could be a promising feature to develop the models on and is worth looking into moving forward.

10 Bibliography

[1] B. M. Nicolaï *et al.*, “Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review,” *Postharvest Biol. Technol.*, vol. 46, no. 2, pp. 99–118, Nov. 2007.

[2] B. Park and R. Lu, *Hyperspectral imaging technology in food and agriculture*. 2015.

[3] A. F. H. Goetz, “Measuring the Earth from Above: 30 years (and Counting) of Hyperspectral Imaging | Features | Jun 2011 | Photonics Spectra,” *Photonics Media*. [Online]. Available: https://www.photonics.com/Article.aspx?AID=47298. [Accessed: 31-May-2019].

[4] A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, “Imaging Spectrometry for Earth Remote Sensing,” *Science (80-.).*, vol. 228, no. 4704, pp. 1147–1153, Jun. 1985.

[5] M. S. Moran, Y. Inoue, and E. M. Barnes, “Opportunities and limitations for image-based remote sensing in precision crop management,” *Remote Sens. Environ.*, vol. 61, no. 3, pp. 319–346, Sep. 1997.

[6] P. Martinsen and P. Schaare, “Measuring soluble solids distribution in kiwifruit using near-infrared imaging spectroscopy,” 1998.

[7] L. T. Evans and I. F. Wardlaw, “Wheat,” *Photoassimilate Distrib. Plants Crop. Source-Sink Relationships*, vol. 60, no. 6, pp. 501–518, 2017.

[8] “File:Naked and hulled wheat.jpg - Wikimedia Commons.” [Online]. Available: https://commons.wikimedia.org/wiki/File:Naked_and_hulled_wheat.jpg. [Accessed: 14-Mar-2019].

[9] K. M. Heinze and M. George, “THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L ’ UNIVERSITÉ DE MONTPELLIER From Grain to Granule : The Biomechanics of Wh eat Grain Fractionation with a Focus on the Role of Starch Granules Valérie LULLIEN-PELLERIN,” 2017.

[10] P. Sáez-Plaza, T. Michałowski, M. J. Navas, A. G. Asuero, and S. Wybraniec, “An Overview of the Kjeldahl Method of Nitrogen Determination. Part I. Early History, Chemistry of the Procedure, and Titrimetric Finish,” *Crit. Rev. Anal. Chem.*, vol. 43, no. 4, pp. 178–223, Oct. 2013.

[11] F. M. Anjum and C. E. Walker, “Review on the significance of starch and protein to wheat kernel hardness,” *J. Sci. Food Agric.*, vol. 56, no. 1, pp. 1–13, 1991.

[12] M. S. Ram, F. E. Dowell, L. Seitz, and G. Lookhart, “Development of standard procedures for a simple, rapid test to determine wheat color class,” *Cereal Chem.*, vol. 79, no. 2, pp. 230–237, 2002.

[13] “Flour Analysis - NDSU Wheat Quality &amp; Carbohydrate Research,” *NORTH DAKOTA STATE UNIVERSITY*. [Online]. Available: https://www.ndsu.edu/faculty/simsek/wheat/flour.html. [Accessed: 21-May-2019].

[14] N. Caporaso, M. B. Whitworth, and I. D. Fisk, “Protein content prediction in single wheat kernels using hyperspectral imaging,” *Food Chem.*, vol. 240, pp. 32–42, 2018.

[15] H. Zhang, B. Gu, J. Mu, P. Ruan, and D. Li, “Wheat Hardness Prediction Research Based on NIR Hyperspectral Analysis Combined with Ant Colony Optimization Algorithm,” *Procedia Eng.*, vol. 174, pp. 648–656, 2017.

[16] S. Q. Yang, D. J. He, and J. F. Ning, “Predicting wheat kernels’ protein content by near infrared

hyperspectral imaging," *Int. J. Agric. Biol. Eng.*, 2016.

[17] N. Gorretta, J. M. Roger, M. Aubert, V. Bellon-Maurel, F. Campan, and P. Roumet, "Determining vitreousness of durum wheat kernels using near infrared hyperspectral imaging," *J. Near Infrared Spectrosc.*, vol. 14, no. 4, pp. 231–239, 2006.

[18] F. Mariotti, D. Tomé, and P. P. Mirand, "Converting nitrogen into protein - Beyond 6.25 and Jones' factors," *Crit. Rev. Food Sci. Nutr.*, vol. 48, no. 2, pp. 177–184, 2008.

[19] R. Owusu-Apenten, *Food Protein Analysis*. CRC Press, 2002.

[20] J. Martín, L. F. Sarria, and A. G. Asuero, "The Kjeldahl Titrimetric Finish: On the Ammonia Titration Trapping in Boric Acid," in *Advances in Titration Techniques*, InTech, 2017.

[21] D. V Guebel, B. C. Nudel, and A. M. Giulietti, "A SIMPLE AND RAPID MICRO-KJELDAHL METHOD FOR TOTAL NITROGEN ANALYSIS," 1991.

[22] M. A. Ferree and R. D. Shannon, "Evaluation of a second derivative UV/visible spectroscopy technique for nitrate and total nitrogen analysis of wastewater samples," *Water Res.*, vol. 35, no. 1, pp. 327–332, 2001.

[23] J. C. Valderrama, "THE SIMULTANEOUS ANALYSIS OF TOTAL NITROGEN AND TOTAL PHOSPHORUS IN NATURAL WATERS," 1981.

[24] Chesapeake Bay (manual), "SECTION D.1 ALKALINE PERSULFATE DIGESTION FOR NITROGEN and PHOSPHORUS, TOTAL and DISSOLVED," 2016. [Online]. Available: https://www.chesapeakebay.net/channel_files/19225/alk-persuldigesttn-tdn-tdp-final_mar2016.pdf. [Accessed: 22-May-2019].

[25] A. H. Simonne, R. R. Eitenmiller, H. A. Mills, E. H. Simonne, and C. P. Cresman, "Could the Dumas Method Replace the Kjeldahl Digestion for Nitrogen and Crude Protein Determinations in Foods?," *J. Sci. Food Agric.*, vol. 73, no. 1, pp. 39–45, 2002.

[26] T. Saint-Denis and J. Goupy, "Optimization of a nitrogen analyser based on the Dumas method," *Anal. Chim. Acta*, vol. 515, no. 1, pp. 191–198, 2004.

[27] M. Thompson, L. Owen, K. Wilkinson, R. Wood, and A. Damant, "A comparison of the Kjeldahl and Dumas methods for the determination of protein in foods, using data from a proficiency testing scheme," *Analyst*, vol. 127, no. 12, pp. 1666–1668, 2002.

[28] P. Sáez-Plaza, M. J. Navas, S. Wybraniec, T. Michałowski, and A. G. Asuero, "An Overview of the Kjeldahl Method of Nitrogen Determination. Part II. Sample Preparation, Working Scale, Instrumental Finish, and Quality Control," *Crit. Rev. Anal. Chem.*, vol. 43, no. 4, pp. 224–272, Oct. 2013.

[29] D. Harvey, "13.4: Flow Injection Analysis - Chemistry LibreTexts," 2016. [Online]. Available: https://chem.libretexts.org/Textbook_Maps/Analytical_Chemistry/Book%3A_Analytical_Chemistry_2.0_(Harvey)/13_Kinetic_Methods/13.4%3A_Flow_Injection_Analysis. [Accessed: 12-Oct-2018].

[30] J. F. C. C. Lima, C. Delerue-Matos, and M. Carmo Vaz, "Flow-injection analysis of Kjeldahl nitrogen in milk and dairy products by potentiometric detection," *Anal. Chim. Acta*, vol. 385, no. 1–3, pp. 437–441, 1999.

[31] J. J. R. Rohwedder and C. Pasquini, “Differential conductimetry in flow injection. Determination of ammonia in Kjeldahl digests,” *Analyst*, vol. 116, no. 8, pp. 841–845, 1991.

[32] X. L. Su, L. H. Nie, and S. Z. Yao, “Determination of ammonium in Kjeldahl digests by gas-diffusion flow-injection analysis with a bulk acoustic wave-impedance sensor,” *Talanta*, vol. 44, no. 11, pp. 2121–2128, 1997.

[33] C. Pasquini and L. C. de Faria, “Flow-injection determination of ammonia in kjeldahl digests by gas diffusion and conductometry,” *Anal. Chim. Acta*, vol. 193, no. C, pp. 19–27, 1987.

[34] L. C. Davis and G. A. Radke, “Measurement of protein using flow injection analysis with bicinchoninic acid,” *Anal. Biochem.*, vol. 161, no. 1, pp. 152–156, 1987.

[35] I. Pasha, F. M. Anjum, and C. F. Morris, “Grain hardness: A major determinant of wheat quality,” *Food Sci. Technol. Int.*, vol. 16, no. 6, pp. 511–522, 2010.

[36] M. Manley, L. van Zyl, and B. G. Osborne, “Deriving a grain hardness calibration for southern and western cape ground wheat samples by means of the particle size index (PSI) method and Fourier transform near infrared (FT-NIR) spectroscopy,” *South African J. Plant Soil*, vol. 18, no. 2, pp. 69–74, 2001.

[37] B. G. Osborne and R. S. Anderssen, “Single-kernel characterization principles and applications,” *Cereal Chem.*, vol. 80, no. 5, pp. 613–622, 2003.

[38] E. B. Maghirang and F. E. Dowell, “Hardness measurement of bulk wheat by single-kernel visible and near-infrared reflectance spectroscopy,” *Cereal Chem.*, vol. 80, no. 3, pp. 316–322, 2003.

[39] M. Başlar, F. Kalkan, M. Kara, and M. Fatih, “Correlation between the protein content and mechanical properties of wheat,” *Turk J Agric*, vol. 36, pp. 601–607, 2012.

[40] A. N. Sieber, T. Würschum, and C. F. H. Longin, “Vitreosity, its stability and relationship to protein content in durum wheat,” *J. Cereal Sci.*, vol. 61, pp. 71–77, 2015.

[41] G. R. Branković, D. Dodig, M. Z. Zorić, G. G. Šurlan-Momirović, V. Dragičević, and N. Durić, “Effects of climatic factors on grain vitreousness stability and heritability in durum wheat,” *Turkish J. Agric. For.*, vol. 38, no. 4, pp. 429–440, 2014.

[42] “Pohl Farinator for Cutting Kernels | Grobacker | Bastak Instruments.” [Online]. Available: https://bastak.com/eng/products/pohl-farinator-for-cutting-kernels. [Accessed: 15-May-2019].

[43] P. Williams and K. Norris, *Near-infrared Technology in the Agricultural and Food Industries*, 2nd ed. American Association of Cereal Chemists, 1987.

[44] Darekk2, “File:Anharmonic oscillator.gif - Wikimedia Commons,” 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:Anharmonic_oscillator.gif. [Accessed: 02-Apr-2019].

[45] “Free practical monographs written by experts | Metrohm,” *Metrohm*. [Online]. Available: https://www.metrohm.com/en/company/news/news-free-monographs/#. [Accessed: 30-Apr-2019].

[46] N3bulous and KYN, “File:Pinhole.svg - Wikipedia.” [Online]. Available: https://en.wikipedia.org/wiki/File:Pinhole.svg. [Accessed: 07-Apr-2019].

[47] KYN, "File:Pinhole2.svg - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/File:Pinhole2.svg. [Accessed: 07-Apr-2019].

[48] Wilhelm Burger • Mark J. Burge, *Principles of digital image processing*. 2009.

[49] L. Ziph-Schatzberg, "Hyperspectral Imaging Enables Industrial Applications | Features | Vision Spectra," *Vision spectra*. [Online]. Available: https://www.photonics.com/Articles/Hyperspectral_Imaging_Enables_Industrial/a56804. [Accessed: 05-Jun-2019].

[50] F. Dell'endice, J. Nieke, B. Koetz, M. E. Schaepman, and K. Itten, "ISPRS Journal of Photogrammetry and Remote Sensing Improving radiometry of imaging spectrometers by using programmable spectral regions of interest," *ISPRS J. Photogramm. Remote Sens.*, vol. 64, pp. 632–639, 2009.

[51] M. Uegly, "File:Sasiedztwa 4 8.svg - Wikimedia Commons." [Online]. Available: https://commons.wikimedia.org/wiki/File:Sasiedztwa_4_8.svg. [Accessed: 14-May-2019].

[52] K. Dunn, "Process Improvement Using Data," 2019.

[53] E. Bulut and S. Kurt, "A performance assessment of model selection criteria when the number of objects is much larger than the number of variables in plsr," vol. 4, no. 6, pp. 257–264, 2012.

[54] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends Anal. Chem.*, vol. 28, no. 10, pp. 1201–1222, 2009.

[55] C. M. Andersen and R. Bro, "Variable selection in regression-a tutorial," *J. Chemom.*, vol. 24, no. 11–12, pp. 728–737, 2010.

[56] H. Li, Y. Liang, Q. Xu, and D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Anal. Chim. Acta*, vol. 648, no. 1, pp. 77–84, Aug. 2009.

[57] M. Hrušková and I. Švec, "Wheat hardness in relation to other quality factors," *Czech J. Food Sci.*, vol. 27, no. 4, pp. 240–248, 2009.

[58] D. Cozzolino, I. Delucchi, M. Kholi, and D. Vázquez, "Use of Near Infrared Reflectance Spectroscopy to Evaluate Quality Characteristics in Whole-Wheat Grain," *Agric. Técnica*, 2009.

[59] L. Breiman, "Random Forests," 2001.

[60] "Create bag of decision trees - MATLAB - MathWorks Nordic." [Online]. Available: https://se.mathworks.com/help/stats/treebagger.html. [Accessed: 06-Jun-2019].

[61] S. Kucheryavskiy, "Chemometrics and Intelligent Laboratory Systems A new approach for discrimination of objects on hyperspectral images ☆," *Chemom. Intell. Lab. Syst.*, vol. 120, pp. 126–135, 2013.

11 Appendix Overview

11.1 Appendix of theoretical interest

11.1.1 Formula for vibrational energy levels for a polyatomic molecule

The formulae of total vibrational energy levels for a polyatomic molecule is:

$$E_v = \sum h \cdot v_r \cdot \left(\boldsymbol{v_r} + \frac{1}{2}\right) + \sum\sum\left(h \cdot x_{rs} \cdot \left(\boldsymbol{v_r} + \frac{1}{2}\right) \cdot \left(\boldsymbol{v_s} + \frac{1}{2}\right)\right) + \cdots$$

$$\left(v = \frac{1}{2 \cdot \pi} \cdot \sqrt{\frac{k}{\mu}}\right), r \leq s$$

Here E_v is the total vibrational energy, h is planks constant, v_r and $\boldsymbol{v_r}$ are respectively the fundamental vibrational frequency and quantum number, r and s are vibrational modes, x_{rs} is a constant of interaction between these two modes.

11.1.2 PCA matrix notation of estimates

Adapted from source [52], the equations and matrix notation for the estimates in PCA looks as:

For one component:

$$x_{i,1}\,(es)' = t_{i,1} \cdot \mathbf{p}_\mathbf{1}'$$

Matrix sizes are:

$$(1\;x\;K) = (1\;x\;1)(1\;x\;K)$$

For two components it translates to:

$$x_{i,2}\,(es)' = t_{i,1} \cdot \mathbf{p}_\mathbf{1}' + t_{i,2} \cdot \mathbf{p}_\mathbf{2}'$$

Matrix sizes are:

$$(1\;x\;K) = (1\;x\;K) + (1\;x\;K)$$

And for all components:

$$x_{i,A}\,(es)' = [\,t_{i,1}\;\; t_{i,2} \ldots\; t_{i,A}\,]\mathrm{P}'$$

$$x_{i,A}\,(es)' = t_i'\mathrm{P}'$$

Matrix sizes are:

$$(1\;x\;K) = (1\;x\;A)(A\;x\;K)$$

Here $x_{i,A}$ is the column vector that contains all predictions of a given i observations variables. Now all observations can be included by matrix notation:

$$X(es) = \mathbf{T} \cdot \mathbf{P}'(= \mathbf{T} \cdot \mathbf{P}^{\boldsymbol{T}})$$

One should note, that in the equation in parenthesis, another sign for transposing of a matrix has been used and the T in italic is not a scores matrix.

Matrix sizes are:

$$(N \; x \; K) = (\; N \; x \; A \;)(\; A \; x \; K \;)$$

Based on the predicted values showcased in the above-mentioned equations and the original real value residual error can be calculated:

$$e_{i},'_{A} = x_i - x(es)'_{i},_{A}$$

Matrix sizes are:

$$(1 \; x \; K) = (1 \; x \; K \;) - (\; 1 \; x \; K \;)$$

Furthermore, when incorporating previous definitions and formulating it in matrix basis, the following structure for the data appears [52]:

$$e_{i},'_{A} = x_i - x(es)'_{i},_{A} = x_i - t'_i \mathrm{P}'$$

$$\leftrightarrow$$

$$x'_i = t'_i \mathbf{P}' + e_{i},_{A}$$

Size: (1 x K) = (1 x A)(A x K) + (1 x K)

In matrix form:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} = \mathbf{X}(\mathbf{es}) + \mathbf{E}$$

Size: (N x K) = (N x A)(A x K) + (N x K)

Here X(es) represents the explained variance and E represents the residual variance.

11.1.3 Principle Component Regression

The underlaying basis starts in the data matrix X, size N x K, where N is objects and K independent variables for each object, as previously stated. Now, present is also the y vector, a vector consisting of the dependent variables, size N x 1. By using the method of PCA it was possible to enter the “PC space” by using the direction vector (the loadings) on the original data to get the scores, placements in the PC space as follows [52]:

$$\mathbf{T} = \mathbf{XP}$$

Now, remember from MLR that:

$$y(es) = \mathbf{X}\mathrm{b}$$

Entering the PC space, this can be translated to:

$$y(es) = \mathbf{T}\mathrm{b}$$

b is found by:

$$\mathrm{b} = (\mathbf{T}'\mathbf{T})^{-\mathbf{1}}\mathbf{T}'\mathrm{y}$$

11.2 FIA: Chemicals

The chemicals used in the whole set-up as stated in the FOSS application note AN 5202 are:

Sulphanilamide (4-aminobenzenesulfonamide), $C_6H_8N_2O_2S$

N-(1-naphtyl)-Ethylene Diamine Dihydrochloride , $C_{12}H_{14}N_2$ x 2 HCl

Hydrochloric acid; HCl , 37%

Sodium Nitrite, $NaNO_2$, dried to constant mass at 150 °C

Sodium Nitrate, $NaNO_3$, dried to constant mass at 105 °C

Ammonium Chloride, NH_4Cl, dried to constant mass at 105 °C

Ammonia, NH_4OH

Sulphuric acid, H_2SO_4

Sodium Hydroxide, NaOH

Potassium peroxodisulphate, $K_2S_2O_8$, Analytical reagent grade.

Boric acid, H_3BO_3

Glycine, H_2NCH_2COOH

Distilled water, H_2O

11.3 DATA from trail experiments

11.3.1 Vitreous Class Summary

This table includes most of the vitreousness measurements for the preliminary experiments of vitreousness. The number in the parentheses is cross-sectional area that is seen as fully vitreous. Only one half of the kernels are displayed here the one that show the clearest vitreousness type.

	1	**2**	**3**
Salut	C	C	C
Alt75	C (45%)	B (10%)	B (30%)
Alt105	B (10)	B (20)	B (20)
AltZH	B (10)	C (5%)	C (5%)
OASIS	C	C (2%)	B (10%)
Sibris	C	B (10%)	B (10%)
Alt325	B (10)	C (5%)	NaN
Pam	C (0%)	C (0%)	C (0%)
Alt530	B (60%)	B (60%)	B (40%)
Altstep	B (20%)	A (80%)	A (90%)

11.3.2 FIA: Sampling and results of preliminary experiments - Part II

OASIS DATA - outliers in red

Weight [mg]	**A**	**B**	**C**
1	11.73	12.84	11.78
2	13.12	14.53	12.26
3	15.49	14.15	13.68
4	9.6	12.08	11.57
5	8.5	8.75	9.01
6	11.93	13.91	12.31
7	10.36	13.32	10.76
8	8.76	11.72	13.08
9	11.7	13.43	10.63
10	8.45	4.38	8.17
11	11.09	12.07	9.22
12	11.86	12.9	12.95
Nitrogen [mg/L]	A	B	C
1	3.07	3.659	3.326
2	3.42	4	3.444
3	4.174	4.247	3.806
4	2.222	3.101	3.07
5	2.292	2.466	2.632
6	3.974	4.705	4.316
7	3	3.357	2.917
8	2.907	3.847	3.953
9	3.142	3.577	3.142
10	1.936	1.168	1.991
11	2.507	2.676	2.314

12	3.435	3.787	4.008
Protein (%)	A	B	C
1	7.629	8.307	8.230
2	7.599	7.934	8.189
3	7.855	8.749	8.110
4	6.747	7.483	7.735
5	7.860	8.215	8.515
6	9.710	9.860	10.220
7	7.060	7.347	7.902
8	9.673	9.568	8.810
9	7.828	7.764	8.616
10	6.679	7.773	7.104
11	6.590	6.463	7.316
12	8.443	8.557	9.022

A75 DATA - outliers in red

Weight [mg]	**A**	**B**	**C**
1	6.7	6.4	6.69
2	7.26	7.2	6.91
3	6.55	7.09	6.38
4	5.36	7.3	5.48
5	6.66	6.48	5.11
6	7.16	7.02	7.27
7	6.04	7.08	5.91
8	7.39	7.15	7.14
9	6.69	7.15	7.14
10	7.21	6.62	5.65
11	5.6	5.59	6.72
12	7.26	7.37	7.39
Nitrogen [mg/L]	A	B	C
1	2.335	2.345	2.296
2	2.416	2.662	2.781
3	2.534	2.69	2.53
4	1.768	2.724	2.19
5	2.033	1.728	1.452
6	2.197	2.338	2.394
7	2.245	2.636	2.223
8	2.305	2.317	2.32
9	2.814	3.021	2.268
10	2.898	2.616	2.073
11	3.569	2.167	2.723
12	2.624	2.465	2.649
Protein (%)	A	B	C
1	10.16	10.68	10.00
2	9.70	10.78	11.73
3	11.28	11.06	11.56
4	9.62	10.88	11.65
5	8.90	7.77	8.28
6	8.94	9.71	9.60
7	10.83	10.85	10.96
8	9.09	9.45	9.47
9	12.26	12.32	9.26
10	11.72	11.52	10.70
11	18.58	11.30	11.81
12	10.54	9.75	10.45

11.3.3 Relative Hardness: Sampling and results of preliminary experiments

In this chapter one can review the measurements of the rupture force preliminary experiments.

Table 29: Rupture force of the OASIS kernels in O1 and O2. Rupture force unit in N.

#	O1	O2
1	147.9	174.8
2	117.7	93.8
3	121.1	87.9
4	93.8	140.1
5	124.5	100.9
6	87.8	101.7
7	78.8	154.4
8	76.8	114.3
9	138.3	160.4
10	96.3	195.9
11	111.3	132.8
12	117.8	93.3
13	110.3	123.3
14	60.3	149.4
15	78.8	143.4
16	101.3	125.8
17	65.3	130.3
18	65.8	132.3
19	80.8	78.3
20	42.7	76.3

Table 30: Rupture force table over the wheat sorts and rupture force measurements for 12 kernels for each sort. Rupture force in N.

	1	2	3	4	5	6	7	8	9	10	11	12
Salut	65	136	160	106	118	145	92	159	91	80	91	77
Oasis	109	121	156	146	135	98	118	92	127	124	91	106
Alt75	65	108	118	122	73	139	89	124	90	148	162	111
Altzh	107	93	127	129	123	96	115	57	120	74	105	114
Altstep	57	54	73	39	82	77	78	91	157	100	63	95
Pam	96	67	132	162	152	128	115	154	140	126	121	73
Sibris	99	58	129	102	109	77	84	163	83	92	150	73
Alt530	143	138	77	109	98	101	108	115	94	69	82	136
Alt325	94	86	109	110	90	81	120	77	101	82	68	88
Alt105	78	65	66	88	95	87	84	85	74	85	153	51

Table 31: Weight and Rupture force for 36 OASIS kernels. Rupture force in N, weight in mg.

#	O1	Weight
1	95.8	38.69
2	52.2	34.45
3	98.3	47.36
4	257	55.92
5	81	48.47
6	143	50.70
7	56	43.17
8	96.1	37.25
9	104.8	52.66
10	58.8	28.28
11	106.6	53.30
12	115.3	31.51
13	86.4	48.65
14	78.3	34.31
15	124.2	50.64
16	102.5	42.12
17	93.9	47.60
18	97.4	45.24
19	96.7	56.87
20	106.4	55.28
21	115.6	42.06
22	64.5	28.24
23	46.9	39.93
24	93	40.26
25	89	43.42
26	57	35.43
27	103	49.76
28	90	30.84
29	118	58.14
30	69	45.54
31	85	41.25
32	54	49.04
33	135	32.19
34	84	32.95
35	81	33.14
36	51	47.50

11.4 Data from main experiments

11.4.1 Stability checks for FIA experiments

The stability checks for each batch of 30 kernels were as follows, *cf. Table 32*.

Table 32: Stability checks for main FIA batch experiments.

	Blank [mg/l]	1.66 mg/l Glycine [mg/l]	5 mg/L NO_3 [mg/l]
Row A B1	0.020	1.885	NA
Row B B1	0.037	1.797	4.957
Row C B1	0.030	1.796	4.940
Row D B1	0.024	1.759	4.888
Row E B1	0.054	1.948	4.986
Row F B1	0.014	1.784	4.998
Row G B1	0.099	1.799	4.945
Row H B1	0.035	1.767	4.938
Row A B2	0.092	1.958	4.975
Row B B2	0.034	2.035	4.937
Row C B2	0.025	1.754	4.925
Row D B2	0.014	1.818	4.951
Row E B2	0.016	1.805	4.996
Row F B2	0.007	1.707	4.920
Row G B2	0.002	1.751	4.921
mean	0.033	1.824	4.948

11.4.2 Multiple comparison plots for main effects

Multiple comparison plot for R-squared CV for the two criteria and the number of principle components, *Figure 51 and Figure 52*..

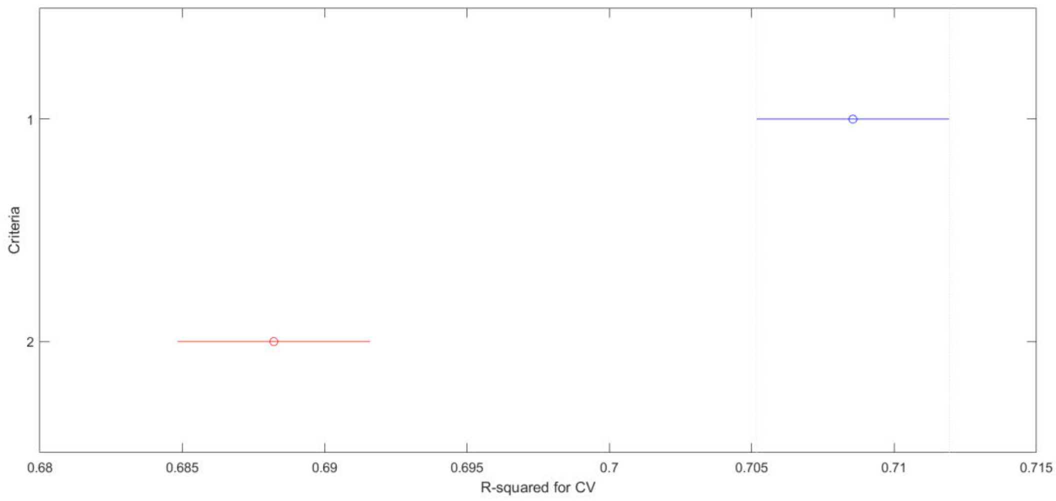


Figure 51: Multiple comparison test on R-squared CV for the Criteria based on protein prediction.

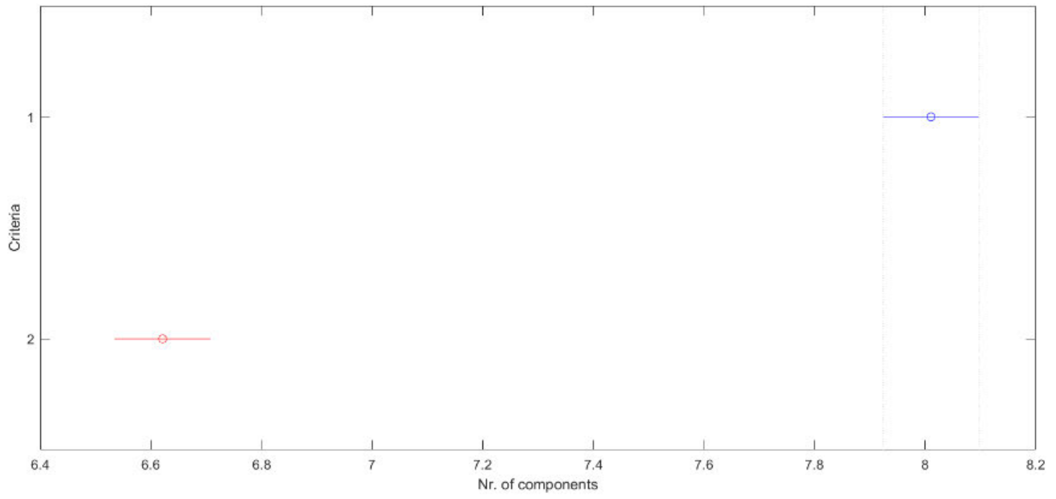


Figure 52: Multiple comparison between the two criteria with regard to nr. of principle components.

Looking at the multiple comparison, it is clear that criteria 1 gives more complex models, *cf. Figure 52*.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 06, 2019

This Agreement between Kronprinsensgade 52 ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4602480413421
License date	Jun 05, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	ISPRS Journal of Photogrammetry and Remote Sensing
Licensed Content Title	Improving radiometry of imaging spectrometers by using programmable spectral regions of interest
Licensed Content Author	Francesco Dell’Endice,Jens Nieke,Benjamin Koetz,Michael E. Schaepman,Klaus Itten
Licensed Content Date	Nov 1, 2009
Licensed Content Volume	64
Licensed Content Issue	6
Licensed Content Pages	8
Start Page	632
End Page	639
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Fig.3
Title of your thesis/dissertation	Development of wheat characterization model based on NIR/HSI for optimization of sorting and quality control of wheat kernels.
Expected completion date	Jun 2019
Estimated size (number of pages)	90
Requestor Location	Kronprinsensgade 52 Kronprinsensgade 52 Esbjerg, 6700 Denmark Attn: Kronprinsensgade 52
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions

apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement

between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript

    - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.9

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 06, 2019

This Agreement between Kronprinsensgade 52 ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4590711281247
License date	May 16, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	TrAC Trends in Analytical Chemistry
Licensed Content Title	Review of the most common pre-processing techniques for near-infrared spectra
Licensed Content Author	Åsmund Rinnan,Frans van den Berg,Søren Balling Engelsen
Licensed Content Date	Nov 1, 2009
Licensed Content Volume	28
Licensed Content Issue	10
Licensed Content Pages	22
Start Page	1201
End Page	1222
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Fig. 14
Title of your thesis/dissertation	Development of wheat characterization model based on NIR/HSI for optimization of sorting and quality control of wheat kernels.
Expected completion date	Jun 2019
Estimated size (number of pages)	90
Requestor Location	Kronprinsensgade 52 Kronprinsensgade 52 Esbjerg, 6700 Denmark Attn: Kronprinsensgade 52
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions

established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of

any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript

  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.9

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jun 06, 2019

This Agreement between Kronprinsensgade 52 ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4590180900945
License date	May 15, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	Journal of Cereal Science
Licensed Content Title	Vitreosity, its stability and relationship to protein content in durum wheat
Licensed Content Author	Alisa-N. Sieber,Tobias Würschum,C. Friedrich H. Longin
Licensed Content Date	Jan 1, 2015
Licensed Content Volume	61
Licensed Content Issue	n/a
Licensed Content Pages	7
Start Page	71
End Page	77
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	Fig 1
Title of your thesis/dissertation	Development of wheat characterization model based on NIR/HSI for optimization of sorting and quality control of wheat kernels.
Expected completion date	Jun 2019
Estimated size (number of pages)	90
Requestor Location	Kronprinsensgade 52 Kronprinsensgade 52 Esbjerg, 6700 Denmark Attn: Kronprinsensgade 52
Publisher Tax ID	GB 494 6272 12
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions

established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.
3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."
4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.
5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.
6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.
7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.
10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).
13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of

any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
    - via their non-commercial person homepage or blog
    - by updating a preprint in arXiv or RePEc with the accepted manuscript

    - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
    - directly by providing copies to their students or to research collaborators for their personal use
    - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
    - via non-commercial hosting platforms such as their institutional repository
    - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.
The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.
If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.
Additional Terms & Conditions applicable to each Creative Commons user license:
CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.
CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.
CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0.
Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.
Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.9

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.