Stefan Slob

Increasing the information content of the annual Finnish Grain Survey using a handheld NIR instrument.

Metropolia University of Applied Sciences

Degree: Bachelor of engineering

Degree Programme: Environmental engineering

Thesis

05/05/2017

Abstract

Author(s) Title Number of Pages Date	Stefan Slob Increasing the information content of the annual Finnish grain quality survey using a handheld NIR instrument. 32 pages + 4 appendices 5 May 2017
Degree	Bachelor of Engineering
Degree Programme	Environmental Engineering
Specialisation option	Water and Waste Management
Instructor(s)	Veli-Matti Taavitsainen, Principal lecturer Frederick Stoddard, Professor of grain technology

Every year a balance is made of Finland's grain production by Evira (the Finnish Food Safety Authority), to determine the quality and availability of grain for use of food, feed, and other uses. Currently this requires a yearly selection of approximately 1500 farms to send a sample of their harvest to Evira's laboratory for analysis using the Kjeldahl method for nitrogen determination and near infrared spectroscopy.

GrainSense Oy, an agritech start-up from Oulu, has developed a handheld near infrared spectroscopy device which they intend to introduce to the Finnish market. This device will help empower farmers but will also provide a vast amount of data for Evira once implemented.

In this thesis, ~500 samples of 2015's grain analyses were measured by a prototype of the device developed by GrainSense Oy to determine if the device can provide adequate results. These measurements were then compared to the data Evira obtained using the Kjeldahl method. This thesis only aimed to prove that the device can measure protein values accurately due to the quality of the samples, which were measured about half a year after Evira did their measurements on them. The results show a promising correlation between the GrainSense device and Evira's measurements.

Keywords	partial least squares regression, Evira Oy, GrainSense Oy, grain measurement, farms in Finland, cereals

Foreword

This thesis explored the possibilities for improvement of the annual grain survey conducted by the Finnish Food Safety Authority Evira. The thesis topic was proposed after the company GrainSense Oy announced that they had developed a handheld NIR spectroscopy device and was commissioned by the University of Helsinki in a joint collaboration. The samples provided by Evira for this study were the samples of 2015's grain survey and consisted of the four main cereal varieties grown in Finland (barley, rye, oat, and wheat).

GrainSense's device has the benefit of being handheld and battery powered, which enables a whole new measurement routine for farmers by giving them the opportunity to make measurements in the field rather than having to send their samples Evira's laboratory. This could mean better results because the samples are measured directly at point of harvest rather than after being transported to Evira's laboratory.

The aim of this study was to use a prototype provided by GrainSense to measure approximately 500 samples out of the 1200 samples of 2015's grain survey data. Then that data was compared to the data provided by Evira to see if the device can provide adequate results. For this thesis, it was only necessary to compare the protein levels of the samples because that is the main element in deciding the purpose of the grain.

I would like to thank both of my supervisors, Veli-Matti Taavistainen and Frederick Stoddard for helping me to understand what it takes to design an experiment and how to conduct solid data analysis. Similarly, I would like to thank the people at GrainSense for their support and for giving me this amazing opportunity to study something I am passionate about.

May 2017,
Stefan Slob

Contents

Figures

Tables

1 Introduction

Cereal production consumes natural resources and provides a livelihood for approximately 50,000 agricultural enterprises in Finland. The demand for high-quality proteins is ever increasing due to population growth. The profitability of those enterprises has been in a decline over the last couple of years, and the Finnish Food Safety Authority, from here on out referred to as Evira, has made it their mission to enhance the quality and competitiveness of Finland’s cereal production. For this to be achievable, Evira started a collaborative project with the University of Helsinki and a start-up company called GrainSense.

The company has developed the first truly handheld device for grain measurement. The device can measure moisture, oils, proteins and carbohydrates from cereals. Because it is a handheld device, it allows for easy on the spot measuring and it should allow the farmer to get more specific data about his/her plot of land. The device will also be able to share its data with a cloud service via which Evira could see in real-time what kind of yield are achieved. With measurements done in the field (or at least on the farm) an added benefit is that there will be less grain damaged, which in return will yield more accurate data.

The purpose of this thesis was to explore the possibilities for Evira once this device would be implemented and to determine if this makes a significant impact on the quality of the data on cereals in Finland. It will provide a background of the main two organisations involved (GrainSense and Evira) and on ascertaining how the measured data from the GrainSense device correlates with the survey data from the grain survey conducted by Evira. Also, it will give information about the theory involved with doing the measurements. The results of the measurement and what possible impact the introduction of this handheld device could mean to farmers of Finland, their cereal production, and Evira’s measurement data will be discussed. It was expected that the measured values of this would be quite different from Evira’s original data because of the time that has passed in between measurements, the samples were measured about half a year later and differences were found in moisture levels.

1.1 Evira

Evira manages, directs and develops the control of products used in the primary production of foods and agriculture in Finland. It works for the Ministry of Agriculture and Forestry and is often consulted as an expert in the sector. Evira is responsible for making risk assessments regarding Finland’s agricultural production. This way Evira protects the rights of Finland’s consumers and contributes to the competitiveness of agriculture and food production.

1.1.1 Agricultural policy

Evira’s agricultural policy is to ensure that factors related to food safety and the environmental impact of production are in line with what the Finnish people expect from their food. Evira conducts chemical food safety studies that are focused on the nutritional content of food products and new production techniques at various stages of production. Evira participates in steering groups for research projects and provides materials created through control and analytical activities, for use in research. As far as possible, it also performs laboratory analyses for research projects.

1.1.2 Communication towards farmers

The results produced by Evira’s researchers provide the information needed for food supervision in Finland. Research conducted by Evira helps determining legislation and standards for farmers to follow and helps with making economic decisions.

1.1.3 Annual grain survey

In 2015, Evira sent out a request for samples to 1850 farms of which 250 were organic farms. This was done for Evira to monitor the grain harvest. The farms were selected based on the farming and horticultural register of the Natural Resources Institute (Luke) using a sampling method. These farms were part of Luke’s yield survey of 6600 farms in total. To make sure information was gathered in every region of Finland regional coverage was taken into consideration when making the selections. Farms with field with a size of less than five hectares were excluded from the sampling.

A total of 1098 conventionally grown samples were received by the closing date from farms of varying sizes, 351 of which were oats, 342 barley, 219 spring wheat, 85 malting barley, 58 rye and 43 samples were winter wheat. A total of 118 samples were received from the organic farms and 60 of these were oats, 13 barley, 11 spring wheat, 5 malting barley, 24 rye and 5 were samples of winter wheat. (Finnish Food Safety Authority Evira, 2017)

On average yearly 1500 farms take part in the survey. (Finnish Food Safety Authority Evira, 2016)

1.2 GrainSense

1.2.1 About the company

GrainSense is a Finnish start-up company based in Oulu which is a spinoff from VTT technical research centre of Finland. The company was founded in 2014 and holds one patent and two patent applications. It has developed the world's first truly hand-held device for grain protein measurement and consequently has secured 1.4 million euro of funding and a development loan from Tekes (the Finnish funding agency for innovation).

GrainSense is trying to give farmers greater control over the quality, cost and pricing of their crops with a cloud-based service offering useful insights to produce their crops and handheld device that can measure key determinants of the harvest value and processing cost of grains. (GrainSense Oy, 2016)

1.2.2 About the device

The device measures near infrared absorbance between 700nm to 1100nm, it is battery-powered and is outfitted with GPS and is capable to store its data directly on to the cloud.

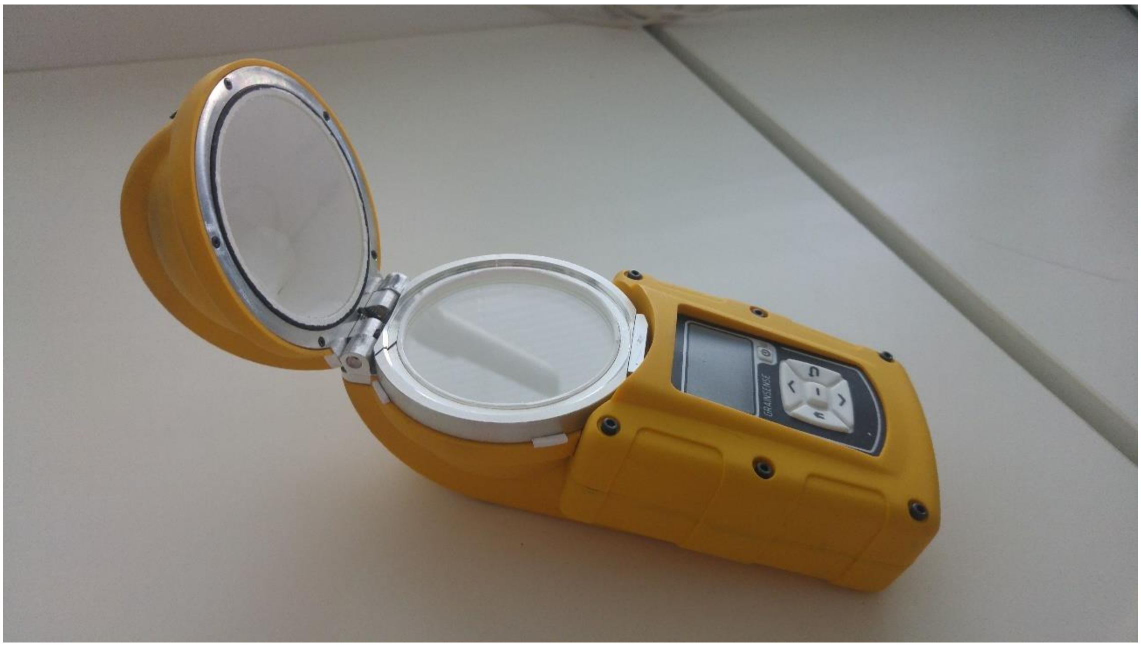

Figure 1. GrainSense device prototype

To make a measurement the farmer needs to open the lid of the device put a handful of kernels onto a dish and close the lid (shown in Figure 1). In both the lid and the device, there are two hemispheres which form an integrated sphere when closed. An integrated sphere (as shown schematically in Figure 2) diffuses the light shined into it. This adds the benefit that the sample does not need to be properly mixed over the dish because the light will average out. After measuring the device will display values like moisture, starch and protein.

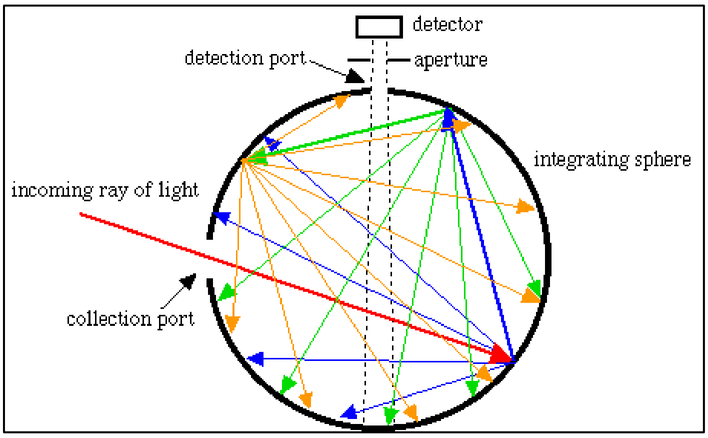

Figure 2. Working principle of an integrated sphere

2 General background

2.1 Cereals and their composition

In principle, all cereals are grown in a similar way, they are annual plants and only produce one harvest during their lifetime. Cereals grow best in a moderate climate. Wheat, rye and barley divide into summer and winter varieties, the winter type requires vernalisation by low temperatures; as a result, they are sown in autumn and mature in early summer. Spring cereals are sensitive to frost temperatures and are sown in springtime and mature in midsummer, they require more irrigation and give lower yields than winter cereals. (Koehler and Wieser, 2013)

Cereals produce dry, one-seeded fruits, called the kernel or grain. The anatomy of cereal grains is uniform: fruit and seed coats (bran) enclose the germ and the endosperm, the latter consisting of the starchy endosperm and the aleurone layer. In oats and barley, the husk is fused together with the fruit coat, and, in wheat and rye the husk can be simply removed by threshing; therefore, they are called *naked grain*.

Table 1. Average composition of the four main cereal grains

(g/100 g)	Wheat	Rye	Barley	Oats
Moisture	12.6	13.6	12.1	13.1
Protein (N × 6.25)	11.3	9.4	11.1	10.8
Lipids	1.8	1.7	2.1	7.2
Available carbohydrates	59.4	60.3	62.7	56.2
Fibre	13.2	13.1	9.7	9.8
Minerals	1.7	1.9	2.3	2.9
(mg/kg)				
Vitamin B 1	4.6	3.7	4.3	6.7
Vitamin B 2	0.9	1.7	1.8	1.7
Nicotinamide	51	18	48	24
Pantothenic acid	12	15	6.8	7.1
Vitamin B 6	2.7	2.3	5.6	9.6
Folic acid 0	0.9	1.4	0.7	0.3
Total tocopherols	41	40	22	18

(Koehler & Wieser, 2013)

In Table 1, you can see the main constituents of the four types of grain that are predominantly grown in Finland. This thesis will address moisture, protein, and carbohydrates because these are the factors that determine the price and application of the grain.

2.1.1 Moisture

When crops are left unharvested, they start to diminish in both quality and quantity due to decay and outside influences (e.g. birds, insects, mould). It is, therefore, important to store the grain at the right time to maximize yields because after harvest the physiological changes within the kernel stop. One of the most critical physiological factors in successful grain storage is the moisture content of the crop. The average moisture content of cereal grains is between 11–14%. High moisture content leads to storage problems because it encourages fungal and insect problems, respiration and germination. However, moisture content in the growing crop is naturally high and only starts to decrease as the crop reaches maturity and the grains are drying. (FAO, 2011)

2.1.2 Protein

The average protein content of cereal grains covers a relatively narrow range 8–11%, variations, however, are quite noticeable. Wheat grains, for instance, may vary from less than 6% to more than 20%. The content depends on the type of cereal, growing conditions (soil, climate, fertilization) with the amount and time of nitrogen fertilization being of great influence. Proteins are distributed over the whole grain, their concentration within each compartment, however, is quite different. The germ and aleurone layer of wheat grains, for instance, contain more than 30% proteins, the starchy endosperm ~13%, and the bran ~7%. Regarding the different proportions of these compartments, most proteins of grains are in the starchy endosperm, which is the source of white flours obtained by milling the grains and sieving. (Koehler and Wieser, 2013)

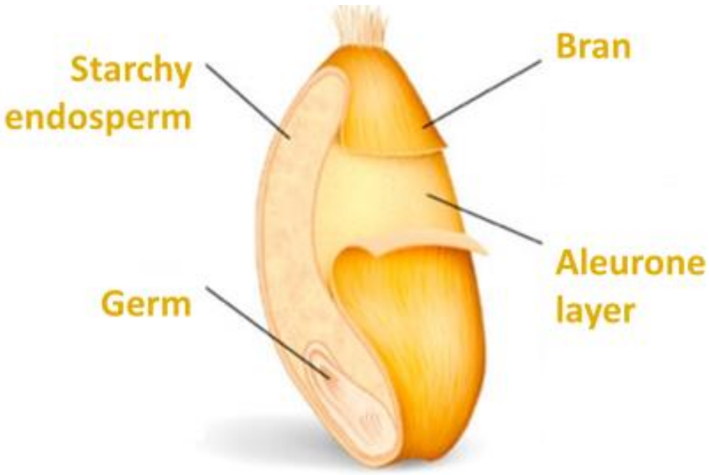

Figure 3. Anatomy of a grain

2.1.3 Carbohydrates

The chemical composition of cereal grains is characterized by the high content of carbohydrates. The available carbohydrates are mainly starch deposited in the endosperm and amount to 56–74% of the grain. Starch is a storage carbohydrate in cereals and an important part of our nutrition. Starch is important for holding water in baked products and is important for the textural properties of many foods, particularly bread and other baked products. Finally, starch is nowadays also an important feedstock for bioethanol or biogas production (Koehler & Wieser, 2013)

2.2 Farms in Finland

2.2.1 Amount of agricultural enterprises

In 2016, there were 50388 agricultural and horticultural enterprises in Finland; this is ~22% less than in 2010, when there were 59483 agricultural enterprises in Finland (shown in Figure 4). This decline happened because only the healthy and viable farms can continue year after year. Those farms often have the means to expand their utilized agricultural area and to mechanize their farms to increase the amount of land a person can manage. Figure 4 shows that farms with a low standard output are declining where farms with a standard output of more than 100000 euro are showing an increase in numbers. This growth suggests that small enterprises ceased their activity as they were incorporated into the bigger ones. (Natural Resources Institute Finland, 2017)

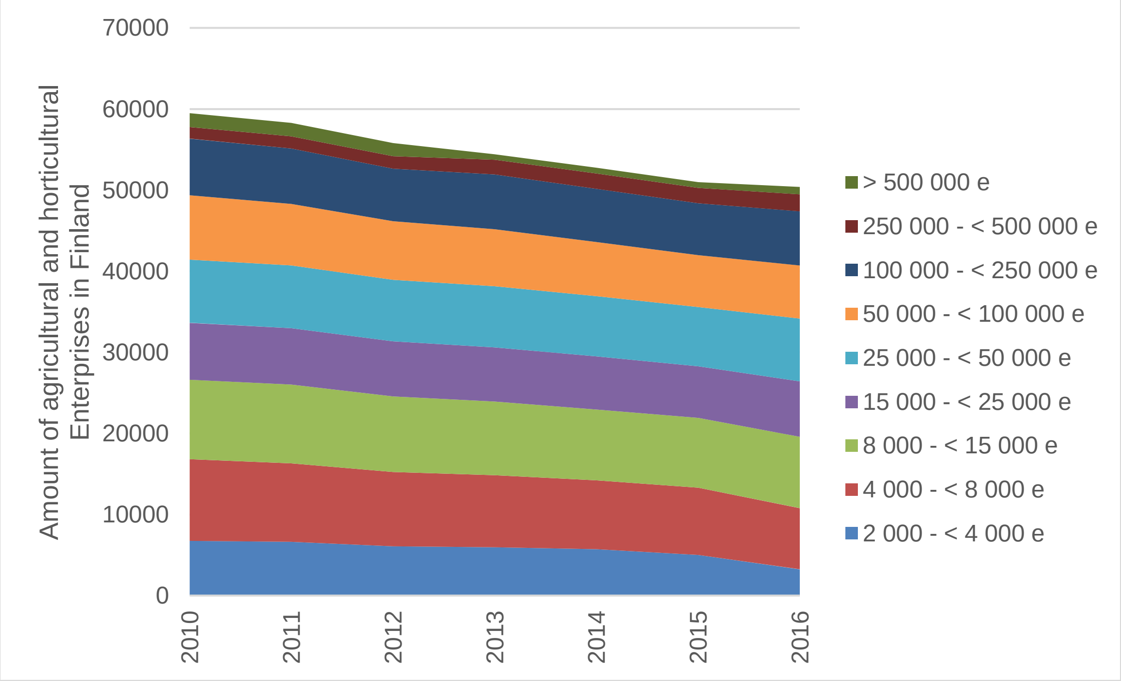

Figure 4. Standard output of Finnish farms per year

For cereal farms, the standard output of an agricultural product is the average monetary value of the agricultural output at farm-gate price in euro per hectare. (Eurostat, 2017)

2.2.2 Utilized farm area in Finland

As shown in Figure 4, farms in Finland has been decreasing for the last two decades by ~22%, but the utilized agricultural area [UAA] has been increasing; the average farm has increased 20% in size in the last 6 years (as shown in Figure 5)

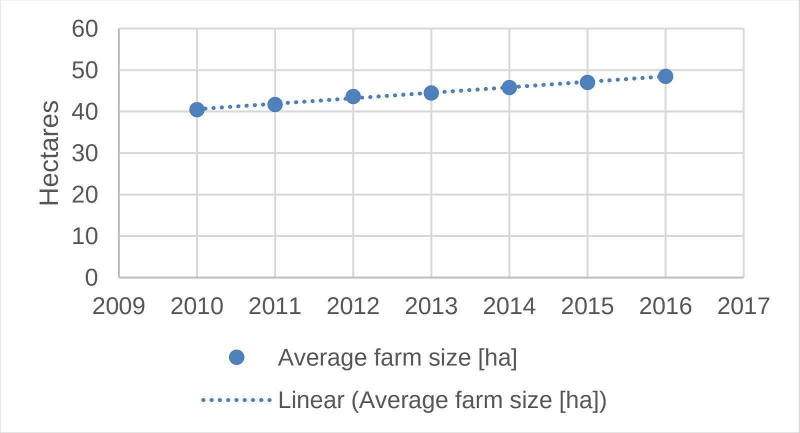

Figure 5. Average size of cereal farms in Finland

This trend does not seem to stop in the near future and implies that the samples received by Evira will be less representative because the samples are taken from bigger fields. (Natural Resources Institute Finland, 2017)

2.2.3 Crop production

Grain crop exceeded four billion kilos in 2014. The amount of four billion kilos has been exceeded on average every other year in the 2000s. In the graph below, you can see that barley is the most popular cultivated grain in Finland, the reason for this is that barley is used for feed stock. Oat shows relatively little change compared to the other grain species and became popular a century ago when horses were used in agriculture, today it is still used as feed for animals. Wheat is mainly used for bread making and has been increasing since 1992. The production of rye has been already for a century; this is mainly because of the cultivation of the other species. (Partela, 2017)

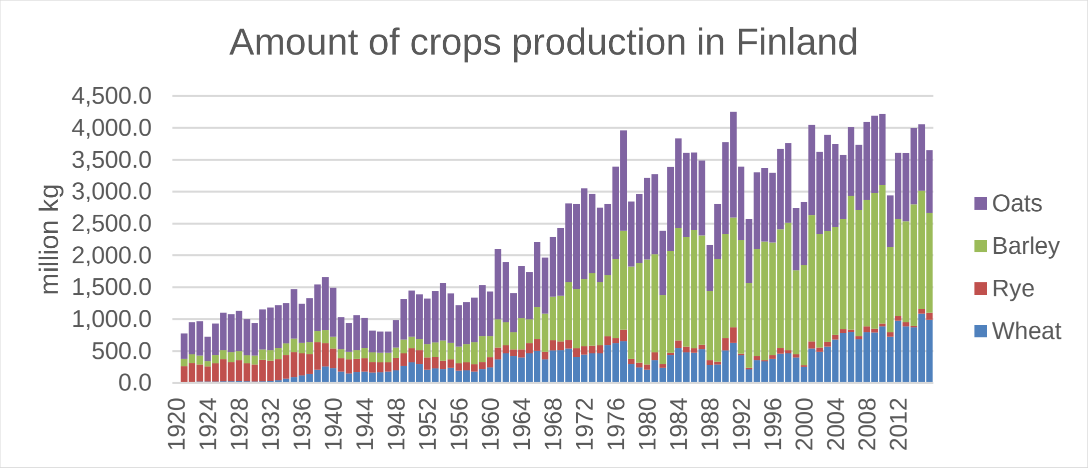

Figure 6. Crop production in Finland

2.2.4 Diverse uses of Cereal grain

After harvesting cereal crops their use gets determined by factors like species, variety, and amount of protein. They will then be sold to the company that has a use for it. The farmer also needs to consider that he needs to have enough grain for next year's harvest. In Figure 7, you see the usage of grain crops over the last decade. A fraction of the grain is used by the farmer's household and a fraction is used for energy production. (Natural Resources Institute Finland, 2016)

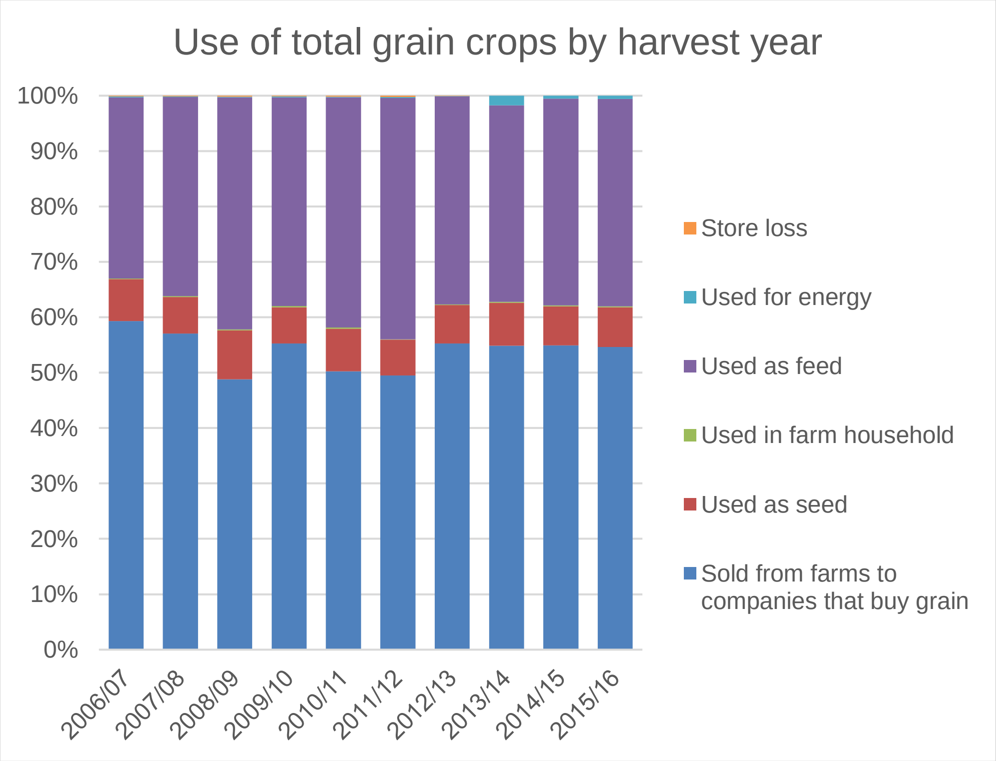

Figure 7. Purpose of grain crops in Finland

The main uses of grain are to feed animals, to feed people, to produce energy, or to be further processed for different kinds of food related products.

2.2.4.1 Food

Each kind of grain has its own area of application; for example, wheat is predominantly used for food because it is the only grain that can store moisture. When looking at food production, the variety of wheat and its protein content play an important role. Wheat varieties have been bred for a wide range of different foods, from biscuits to spaghetti.

2.2.4.2 Feed

Depending on the purpose of the livestock (dairy production or meat production) the feed preparation gets tailored for each stage of the animal's development. Meat, just like most food, has different grades of quality that get sold for different prices. The farmer needs to know about the protein content, the digestible carbohydrate, lipid (oil) content and fibre content. These are all necessary to ensure that the final product meets the desired nutritional quality. In Finland, the main species for animal feed are barley and oat. (Batey, 2017)

2.2.4.3 Industrial

The amount of grain utilised for neither food nor feed purposes has grown in recent years. The main industrial use of grain is to isolate the starch component and then process it further. As much as 60 to 80% of the dry matter of most cereal grains is starch and it is isolated industrially from wheat. The resulting starch may be utilised as it is or it may be processed further. Industrial processing of grains may provide products for human consumption, perhaps as alcoholic beverages or as starches added to foods to give desirable functional properties in the food. (Batey, 2017)

2.2.5 GHG (Greenhouse gas) of cereal production

It is good to realise that the application of fertilizer is a significant contributor to greenhouse gas emissions. In Finland agriculture contributes for ~6,5 million tonnes of CO2 equivalent yearly (Greenhouse gas inventory unit. Statistics Finland , 2016)

A study done by the university of Helsinki (Rajaniemi, Mikkola, & Ahokas, 2011) shows that improving the harvest yield has a strong impact on emissions per kilogram. They found that If the grain yield increased by 20%, the amount of GHG emissions per produced grain kilos decreased by 23%. If the grain yield decreased by 20%, the GHG emissions per produced grain kilos increased by 16%.

3 Analytical methods

3.1 Kjeldahl method

Nitrogen is one of the main constituents for organic materials like protein; therefore, measuring nitrogen in an organic sample can teach us a lot about its content. In the food industry, the Kjeldahl method is universally used as the standard method to determine protein because of its precision and reproducibility. After the nitrogen content of the sample is measured, it can then be converted to crude protein content with the use of a multiplication factor as shown in table 2. (Blamire, 2003)

Table 2. Conversion factor of nitrogen to protein in cereal grain

Commodity	Conversion factor (Nitrogen to protein)
Common wheat	5.7
Durum wheat	5.7
Wheat milling products	5.7 or 6.25
Wheat for feed	6.25
Barley	6.25
Oats	5.7 or 6.25
Rye	5.7

The apparatuses necessary for doing a Kjeldahl nitrogen determination are:

- Mechanical grinder.
- Sieve, with aperture size 0.8 mm.
- Analytical balance, capable of weighing to the nearest 0.001 g.
- Digestion, distillation and titration apparatus.
- A heater

The procedure can be described briefly as follows:

1. The sample is first digested in strong sulfuric acid in the presence of a catalyst, which helps in the conversion of the amine nitrogen to ammonium ions.
2. The ammonium ions are then converted into ammonia gas, heated and distilled. The ammonia gas is led into a trapping solution where it dissolves and becomes an ammonium ion once again.
3. The amount of the ammonia that has been trapped is determined by titration with a standard solution, and a calculation made.
(Blamire, 2003)

3.1.1 Preparation

A representative sample gets sent to the laboratory. It should not have been damaged or changed during transport or storage. The test sample of grain is measured to be at least 200 grams; those 200 grams get grinded until they can pass through the sieve entirely. After thorough mixing a subsample of between 0.5 and 1 gram (rounded until the nearest 0.001g) is taken from the grounded sample. A portion of the remaining grounded sample is then used to determine the moisture level needed for later calculations. (ISO, 2006)

A blank test also needs to be performed to compare the results; this test follows the same procedure but without the sample.

3.1.2 Digestion

The sample is placed in a digestion flask, and 20 ml of sulfuric acid (H_2SO_4, 0.05 mol/l.) is added. 10g of potassium sulphate is added to elevate the boiling point of the sulfuric acid and the combination of 0,30g of titanium oxide and 0,30g of copper(II) sulphate pentahydrate is added as a catalyst.

This total mixture then gets heated to 420 (± 10) °C. After a minimum of 120 minutes of digestion the mixture is left to cool, this is measured from the time that the mixture reached 420 (± 10) °C after being put of the heater. For safety reasons, it is necessary to do this part of the test under a well-ventilated fume hood. (ISO, 2006)

The reaction can be expressed as follows:

$$Sample\ +\ H_2SO_4 \rightarrow\ (NH_4)2SO_4(aq)\ +\ CO_2(g)\ +\ SO_2(g)\ +\ H_2O(g)$$

The result is an ammonium sulphate solution.

3.1.3 Distillation

To distil the mixture, its pH needs to be raised; this is done with sodium hydroxide and has the effect of changing the ammonium (NH_4^+) ions which are dissolved in the liquid, to ammonia NH_3(g):

$$(NH_4)2SO_4(aq)\ +\ 2NaOH\ \rightarrow\ Na_2SO_4(aq)\ + 2H_2O(l)\ +\ 2NH_3(g)$$

That gas is separated away from the mixture by distilling the ammonia by converting it to a volatile gas (by raising the temperature to boiling point) and then trapping the vapours in a trapping solution of boric acid (H_3BO_3). The ammonia is bound to the boric acid in the form of ammonium borate complex:

$$H_3BO_3 + NH_3 \rightarrow NH_4^+ + H_2BO_3$$

3.1.4 Titration

The quantities of acid, and therefore ammonia are determined by adding an indicator dye to the acid/ammonia trapping solution. This dye should turn a strong colour, indicating that a significant amount of the original trapping acid is still present.

By slowly adding small amounts of the sodium hydroxide solution to the acid solution with the dye, it is possible to indicate the "endpoint" has been reached and that all the acid has been neutralized by the base. For the calculations, it is necessary to mark down the volume of the neutralizing base (sodium hydroxide solution) that was necessary to reach the endpoint. (Blamire, 2003)

3.1.5 Calculation

$$w_N = \frac{(V_1 - V_0)T\times 0{,}014\times 100}{m}\times\frac{100}{100 - w_H} = \frac{140T(V_1 - V_0)}{m(100 - w_H)}$$

where:

V_0 = the volume, in millilitres, of the sulfuric acid solution needed for the blank test

V_1 = the volume, in millilitres, of the sulfuric acid solution needed for the test portion

0.014 = the value, in grams, of the quantity of nitrogen equivalent to the use of 1 ml of a 0.5 mol/l sulfuric acid solution

T = the normality of the sulfuric acid solution used for the titration

m = the mass, in grams, of the test portion

w_H: the moisture content

(ISO, 2006)

3.2 Near infrared spectroscopy [NIR]

Near infrared spectroscopy refers to the spectrum of light directly next to the visible spectrum. This ranges from between 750 and 2500 nm in wavelength as shown in Figure 8. Most organic materials have well-defined absorbance and transmittance features at these wavelengths. When infrared light is shined on an organic sample the molecules begin to vibrate, this happens because of the energy inserted to them by the infrared light. This does not happen equally for all molecules, and it is observed that bonds with hydrogen (C-H, N-H, O-H and S-H bonds) show the largest vibrations because hydrogen is the lightest atom and therefore will stand out when measuring an organic sample. (Metrohm AG, 2013)

When a sample is measured using near infrared spectroscopy the output will show a level of absorbance at each measured point across the wavelength. Graphically, this is represented as a wave with absorption peaks showing at the varying wavelengths (this can be seen in Figure 14). These peaks are unique to the chemical composition of the sample and serve as a ‘fingerprint’ for that sample. The reason why this is the case is because the light reflected to the sensor will be less intense because it has lost the energy that went into getting the molecules to vibrate. (Metrohm AG, 2013)

Near infrared spectroscopy can, therefore, accurately measure organic samples and is only limited by the fact that it cannot interpret the chemical composition of the sample without being calibrated. To do this, we first need to link the obtained data with values obtained from the sample using a chemical method. A popular method for this is mentioned above and is called the Kjeldahl method, it is utilised by Evira when measuring grain samples.

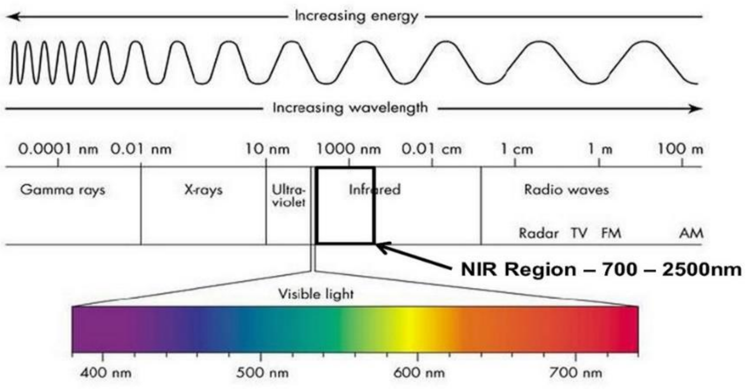

Figure 8. Wavelength region of near infrared

3.2.1 Absorbance

The device GrainSense has built measures the absorbance of the samples. The absorbance is equal to the difference between the logarithms of the intensity of the light entering the sample (I_0) and the intensity of the light transmitted back (I) by the sample:

$$A = log\, I_0 - log\, I = log\,(I_0/I)$$

Due to this absorbance is dimensionless. (Stuart, 2004)

3.2.2 Transmittance

For solid samples, the term transmittance is used in spectrophotometric analysis. It is the ratio between the intensities of light measured with and without the samples. The device Evira uses to measure its samples measures using near infrared transmittance rather than absorbance. (Stuart, 2004)

Transmittance is defined as follows:

$$T = I/I_0$$

and percentage transmittance as follows:

$$\%T = 100 \times T$$

Therefore, absorbance can be expressed as follows:

$$A = -log\,(I/I_0) = -log\, T$$

4 Statistical methods

4.1 Partial least squares regression [PLS]

When working with wavelength data like data obtained with near infrared spectroscopy it is common to use partial least square regression for the construction of a predictive model. This is because wavelength data from for example grain samples will show a high degree of collinearity. PLS compares that wavelength data with component amounts (in my case protein levels) and looks for hidden and underlying relationships between variables and tries to extract those from the data. This way it is possible to create a solid model without overfitting because it might just be that out of the entire wavelength only a couple measuring points account for most the variation. This approach gives partial least square regression an advantage over other methods like multiple linear regression, because even though MLR can also be used with many factors it might happen that the number of factors gets too large (for example, if there are more factors than observations). What would happen in that situation is that you make a model that perfectly fits the sample data but will not be able to predict new data properly (i.e. the model is overfitted).

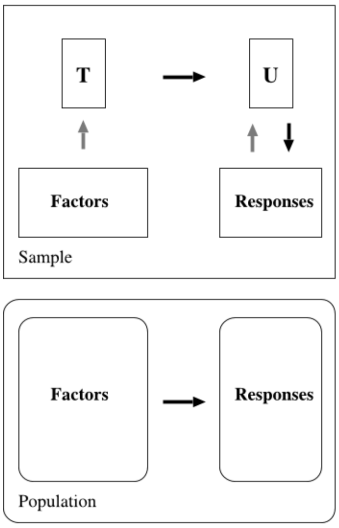

Figure 9. Schematic outline of partial least square regression

Figure 9 gives a schematic outline of the method. As you can see the goal is to use the factors (spectral data) to predict the responses in the population (e.g. protein data). This is achieved indirectly by extracting hidden variables T and U from sampled factors and responses. The extracted factors T are used to predict U responses (also referred to as X-scores and Y-scores). Then the predicted Y-scores are used to construct predictions for the responses. (Tobias, 1995)

5 Design of experiment

5.1 Sample Selection

For my thesis Evira provided a list with protein, starch and moisture levels of the grain samples of 2015’s harvest. For the experiment, a total of 500 samples were selected from the ~1200 samples taken. The set was selected semi-randomly due to availability. I made my initial selection using the provided list and selected a proper range of values with emphasis on the outliers (highest values and lowest values) By the time I got to collect the samples some of the ones were not available anymore. Evira had used a portion of all the samples for different experiments already and I got to collect the samples in September 2016. It was then decided that because my sampling size was so big random selection would still show a proper spread in values ranging from low to high protein, starch and moisture. The division between the 500 grain samples is ~50 rye, ~150 barley, ~150 wheat, and ~150 oat samples.

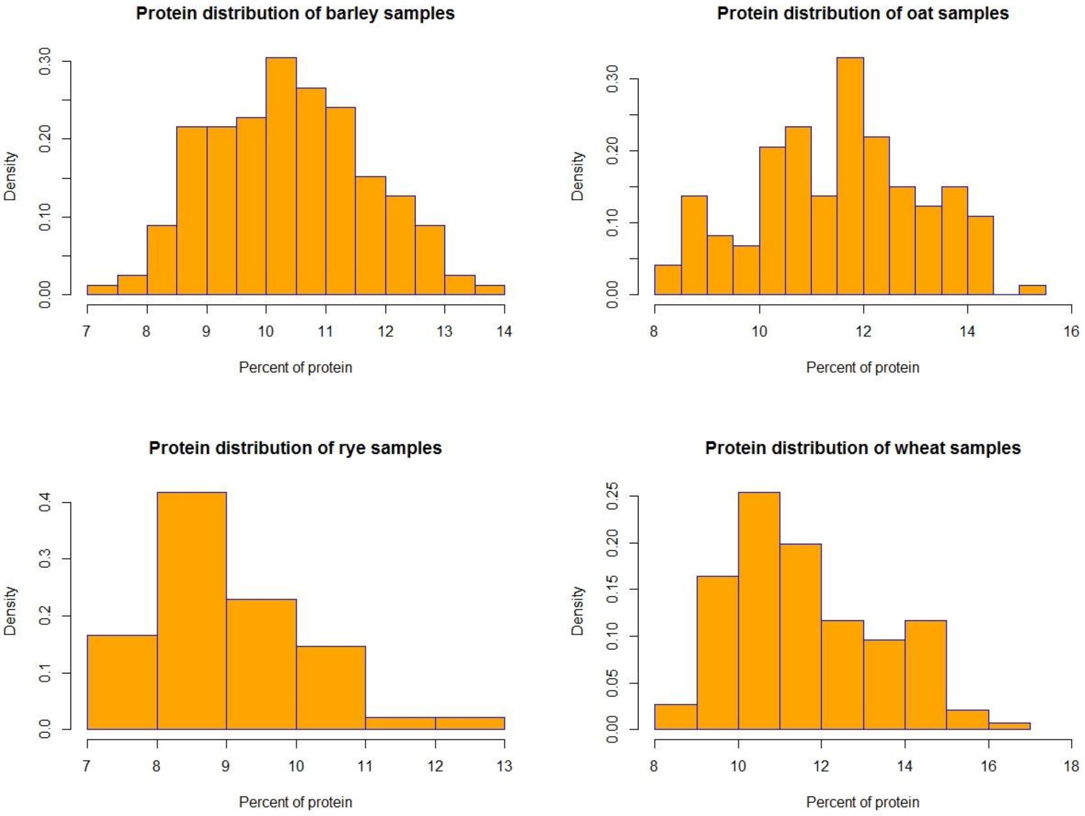

Figure 10. Protein distribution of samples

In Figure 10 you see four histograms of the samples selected sorted by species. All of them show quite a good distribution of protein levels and barley shows the best distribution. In table 1 it showed that the average protein amount of barley, oat, rye and wheat were 11.1%, 10.8%, 9.4% and 11.3% respectively. The average protein of the samples are 10.4%, 11.6%, 9.0% and 11.6%.

5.2 Test setup

Ultimately after selecting the samples I ended up with 158 barley samples, 146 oat samples, 48 rye samples and 146 wheat samples (so 498 in total). From those samples 50 kernels (with the skin intact) were taken. Damaged skin would cause an offset in the amount of carbohydrates measured because the starch would then be detected directly for it has no hull surrounding it anymore. All samples were manually inspected by me when I took selected the 50 kernels. From all the samples the device took 4 replicate measurements to reduce the measuring error.

GrainSense’s device measures the transmittance of the sample with this formula:

$$\mathrm{T}_{(\lambda)} = \frac{P_{sample}(\lambda)}{P_{empty}(\lambda)}$$

where:

$\mathrm{T} = \ transmittance\ measured\ in\ \lambda.$

$P_{sample} = Power\ of\ the\ light\ going\ through\ the\ sample\ measured\ in\ \lambda.$

$P_{empty}\ = Power\ of\ the\ light\ going\ through\ the\ empty\ sample\ dish\ measured\ in\ \lambda.$

The output of the device would be absorbance.

$Absorbance = \ -\log_{10}(\mathrm{T})$

For every measurement, the sample dish (as seen in Figure 11) was emptied and cleaned. Then a reference measurement was taken with no grain in the device, this way it could be assumed that all measurements were independent from each other. For a measurement 50 kernels of a sample were evenly over the dish while leaving some space between them, this way the light shined on them would be able to reach everywhere and this was supposed to produce better results.

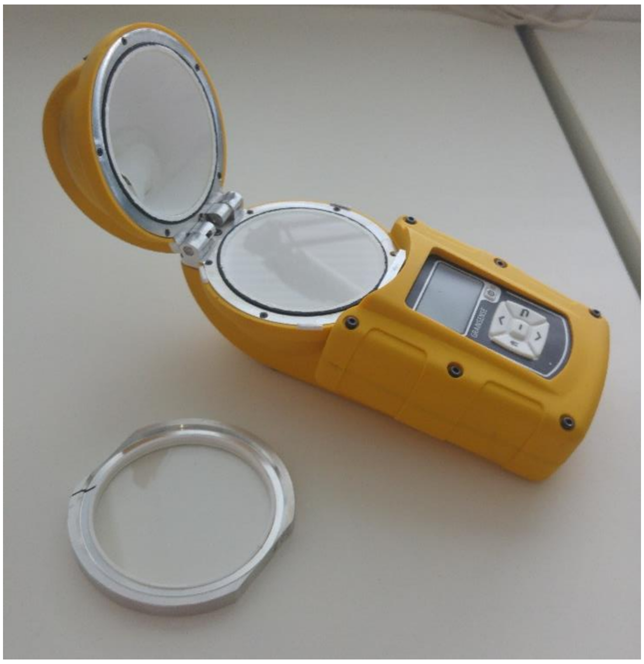

Figure 11. Sample dish

5.3 Data collection

To collect the data, the GrainSense device got connected to a laptop (as seen in Figure 12) from which I could operate the device. This way It was possible to save the results to a text file for later processing. It was also possible this way to let the device do 4 replicate measurements in a row. I saved the measured data from each variety in its own file to make analyses easier later on.

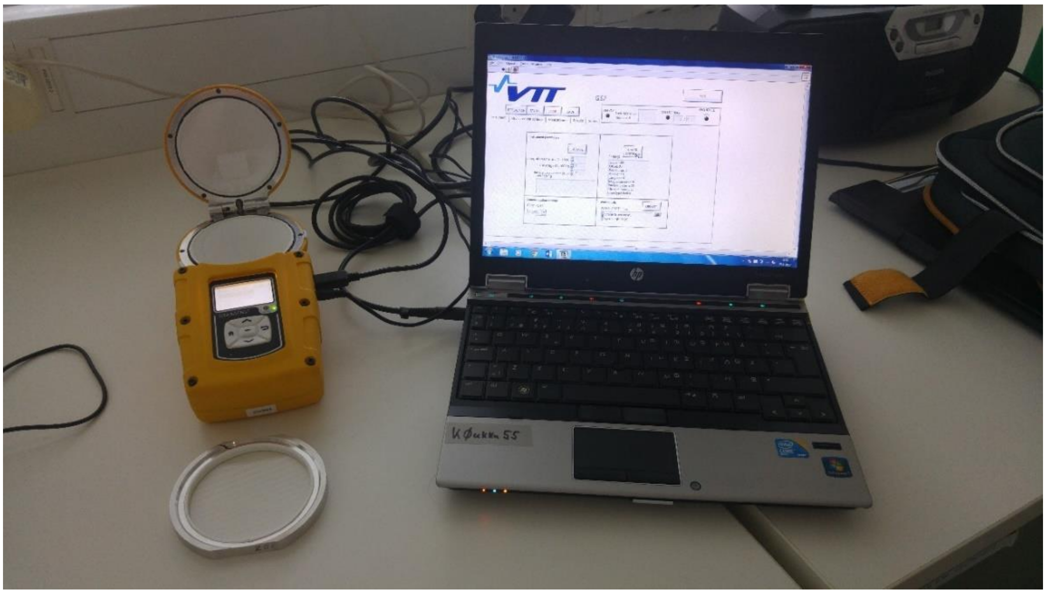

Figure 12. GrainSense device linked to a laptop

After doing all the measurements I made a file for each variety that had a column for sample number and then 76 columns for the measured points along the wavelength. These files provide the factors in my pls model. For the response variables, I again made a text file for each variety, this time the columns were sample number and protein value.

6 Analyses

6.1 Fitting the models

Because of the nature of the measurements the approach to making the model was relatively simple, first a script was written in R for one variety and then it was altered so that same script would work with the other varieties. This way changing things in the script only meant changing some values and titles for different sample sets.

To make a good model the first thing I did was to remove the outliers using a principal component analysis. Principal component analysis finds a new coordinate system in which every measurement has a new (x, y) value. The axes on the PCA plot don't mean anything physical; they are the “principal components" that are chosen to give one of the axes maximal variance.

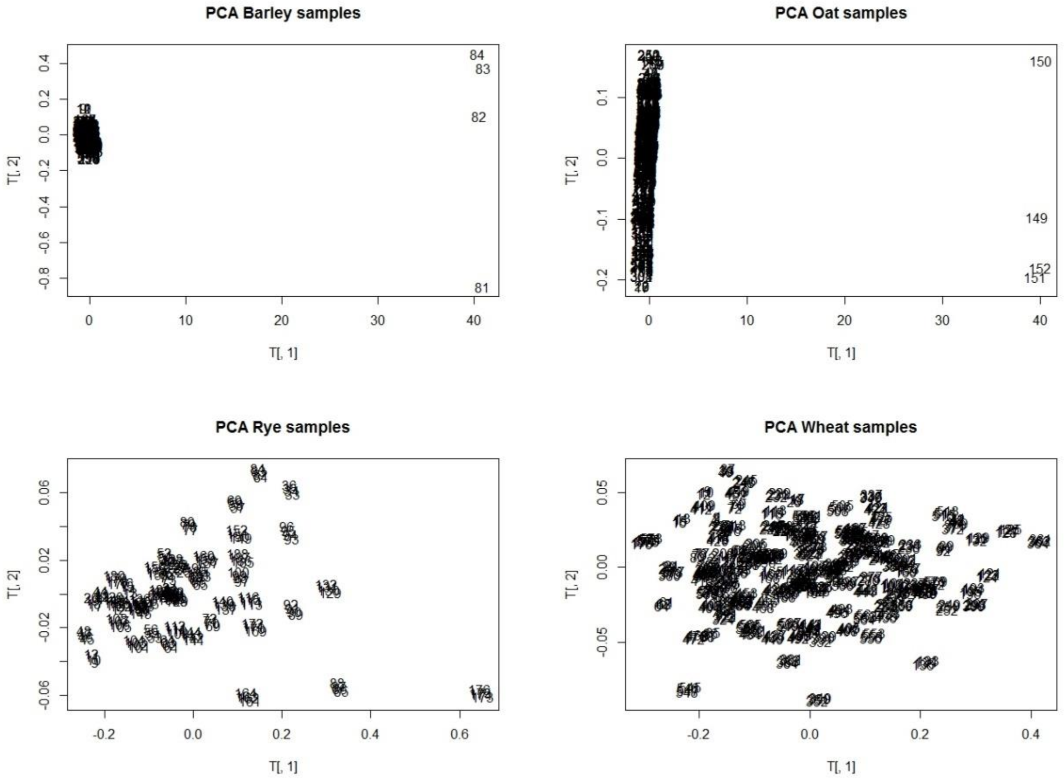

Figure 13. Principal component analysis samples

In Figure 13 you can see the principle component analysis of the four varieties. In barley, oat and rye you can see a that there is an outlier, for wheat it was not needed to remove an outlier.

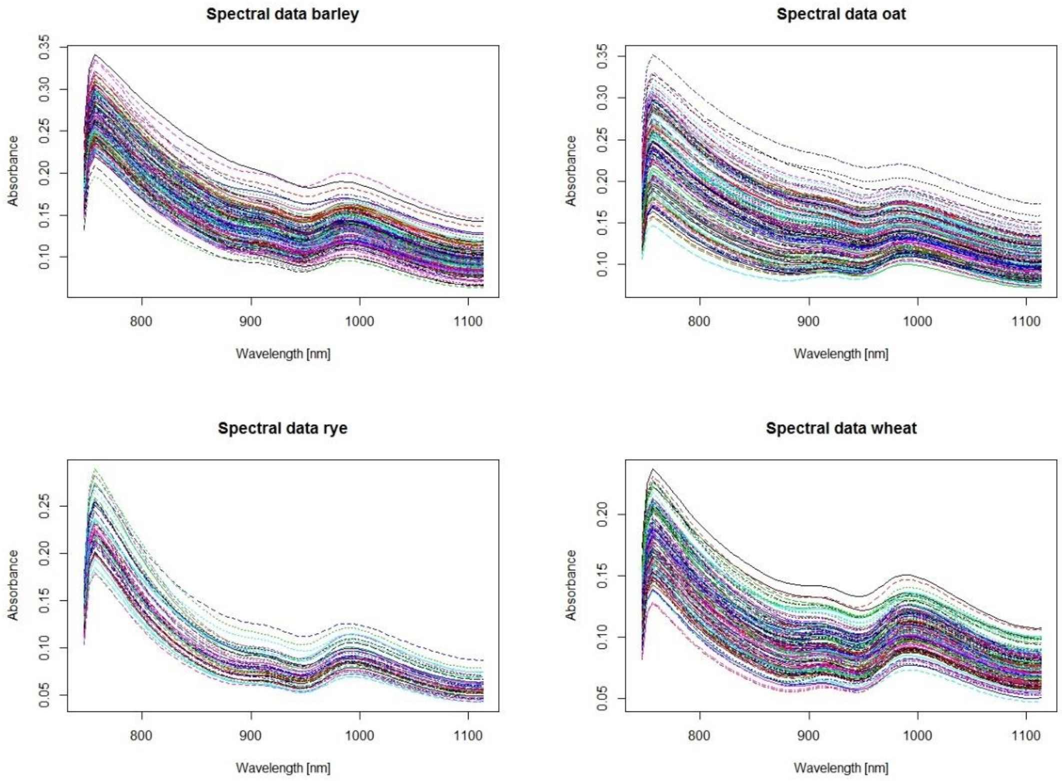

Figure 14. Original spectral data

From Figure 14 you can see the plotted spectral data of the 4 varieties, already from the plots some differences are visible, but these differences cannot be interpreted without making an PLS model to extract the hidden relationships between the spectral data and the protein data. It also becomes clear that rye has a lot less spectral data available as the other varieties.

After removing the outliers, a PLS model was made for each variety to see how well the measured data could be fitted with the protein data. For this task, R’s “PLS” package and I selected cross validation as validation method, this means that a train set and a test set were created from the measured data and that the model was trained with the train set and tested with the test set. Cross validation enabled me to choose an optimal number of dimensions for the model, which was chosen using cross validation RMSEP (Root Mean Square Error of Prediction). The aim was to select the number of dimensions which show the lowest amount of RMSEP. In Figure 15 you can see a plot of the RMSEP of each model.

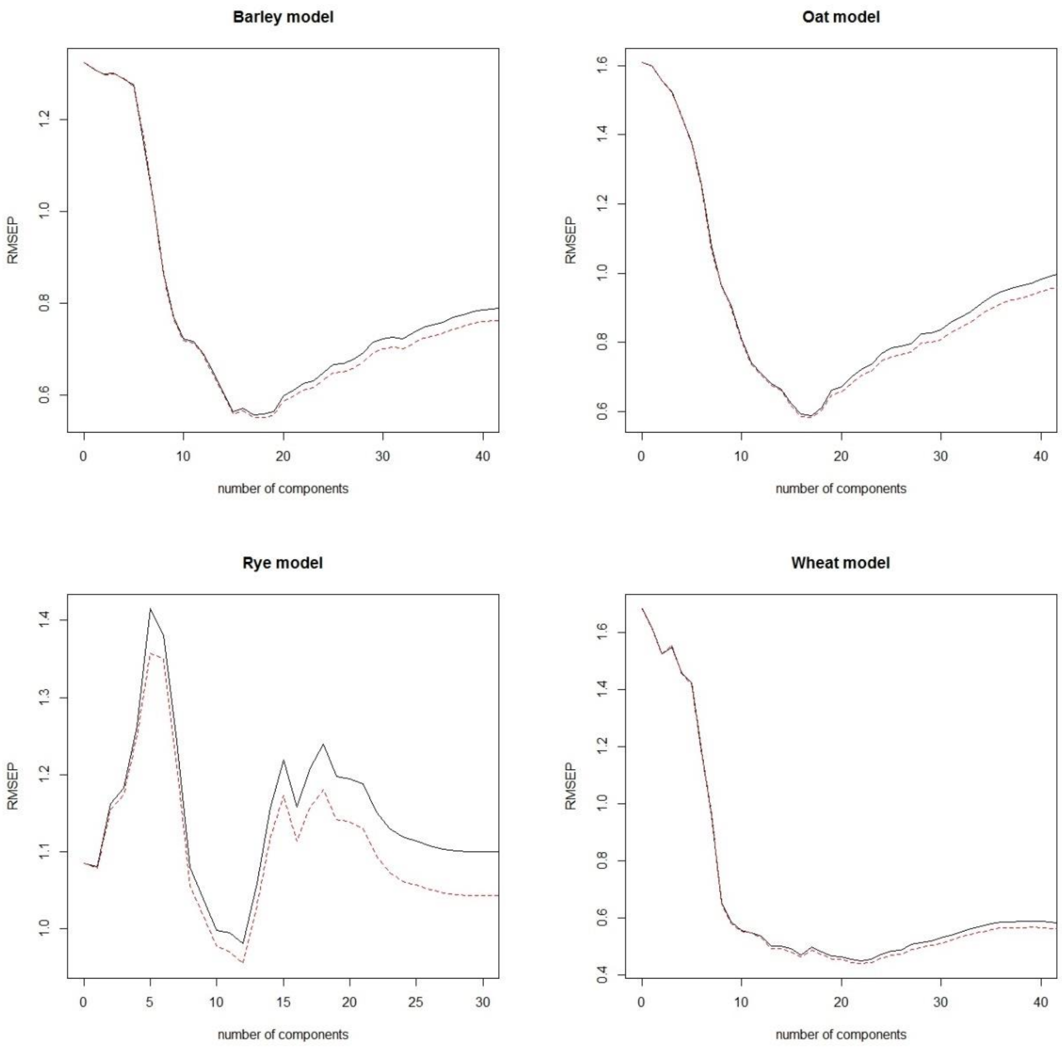

Figure 15. RMSEP of CV

The number of components were different for each model, barley needed 18 dimensions, oat 16, rye 11 and wheat 21.

In Figure 16 you can see the fit of the four models. The blue lines show a 1:1 ratio line which would describe a perfect fit. For several reasons the models did not achieved this perfect fit, but this will be addressed in the discussion section of this thesis. R-squared (R^2) is mentioned in the title, this is the statistical measure of how the measured data correspond with the fitted data. The worst fit is found in the rye model, this probably was due to the small sample size (~50 samples versus ~150 samples in the other species), the age of the sample, and the low sample size. For the other species, a model fit of over 90% was achieved, which given the age of the samples is okay.

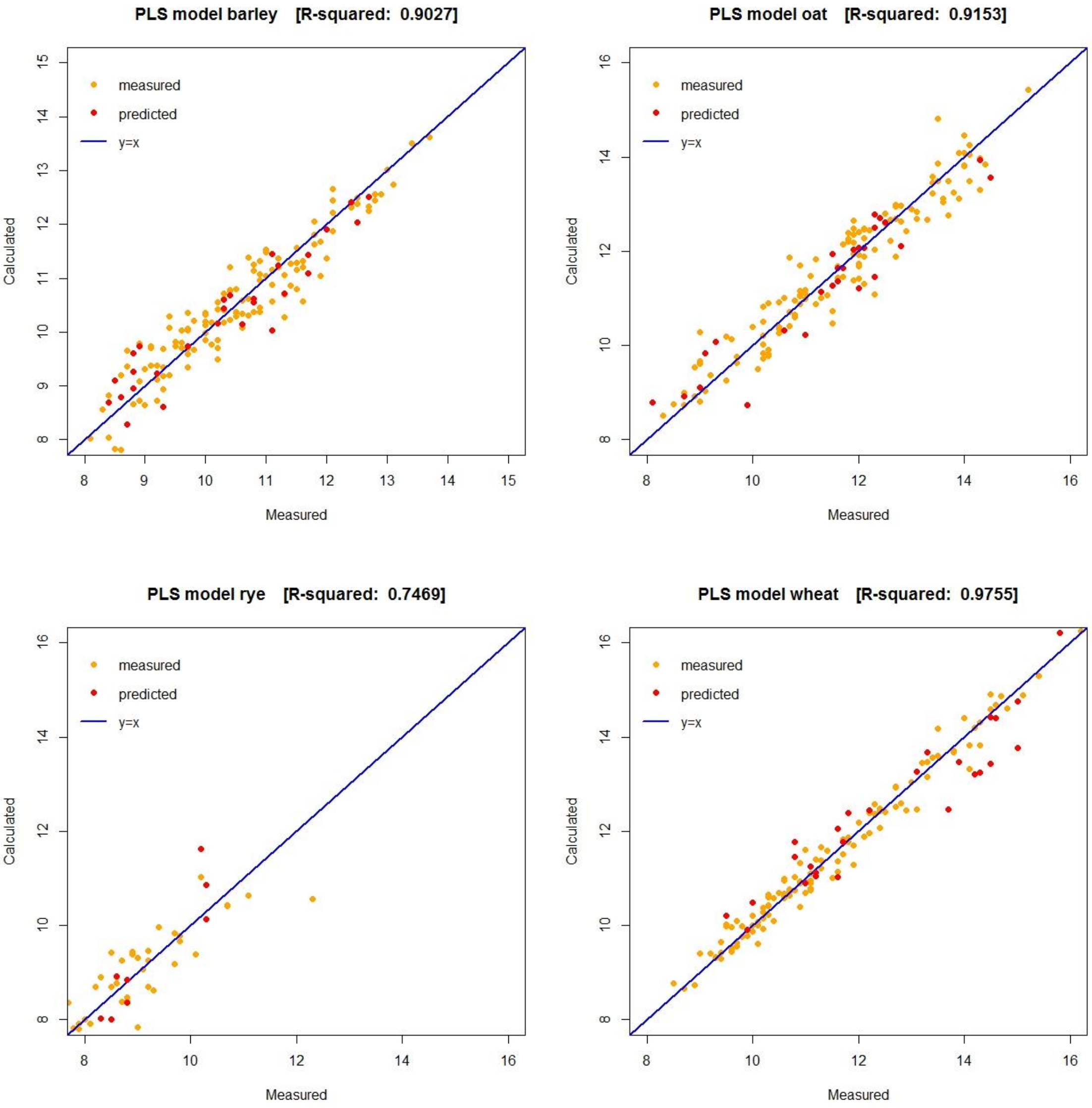

Figure 16. Model fit of the four species

6.2 Degree of correlation

As earlier mentioned partial least square regression can fit models using data that portrays high levels of collinearity (like spectral data) to a response variable (like protein values) but it what would be even better to know how much the results deviated from their actual values. This is achieved by dividing the root mean square by the mean of y. For barley, the deviation was ±4.2%, for oat ±4.5%, for rye ±6.7% and for wheat ±5.3% from Evira's measured data. It shows that the hulled grains show a lower deviation than the grains with the hull fused to the kernel.

7 Discussion

7.1 Comparison with laboratory equipment used by Evira

Evira currently uses a FOSS Infratec™ NOVA device for their measurements, the product datasheet of the FOSS Infratec™ NOVA a maximum of 0.1% variation for measurement of protein in wheat. The Kjeldahl method at Evira has a standard uncertainty of measurement for rye and wheat of ±0,3 % units and for oats and barley ±0,4 % units. This is quite a contrast with the measured data (barley ±4.2%, oat ±4.5%, rye ±6.7% and wheat ±5.3%). However, I do not believe that this is the fault of the device, as shown in Figure 13 the replicates of the measurements are very close to each other, indicating that the device is precise and that the cause of the deviation should originates from other sources.

7.2 Sources of error

7.2.1 Random error

The samples are from 2015's harvest were collected in September 2016; this was the soonest moment it was available to do so because that was when Evira was finished with their measurements on them. Due to the time spent in storage the samples lost moisture and this changed the composition of the grain Evira measured using the Kjeldahl method. This change manifested itself in the form of a different volume percentage of protein, when moisture was taken out of the total composition the percentage of protein on the total increased, while it stayed the same in weight. It was also assumed that the moisture loss would be the same for every sample because of similar storage conditions but this was not tested. Because of the storage conditions being similar for all samples this deviation can be considered somewhat systematic and the PLS model corrects for systematic error. This means that the difference in moisture might not have been the main cause of the deviation.

For rye, a bigger sample set would have helped overcoming the random error and consequently producing a better model, but there wasn't any more rye available at the time. This is because rye is not equally represented in the grain survey, because only a small portion of farms in Finland produce rye compared to the farms that produce barley, oat and wheat.

Another source of error was the individual differences between the grain, because the sample size was so small the difference between individual kernels already provided a deviation from the true value.

7.2.2 Systematic error

All sample measurements were independent from each other; this is because of the blank measurement taken between each sample. There was also no environmental error because all measurements were done in a laboratory at the University of Helsinki.

In hindsight, it would have been more accurate if the samples were weighed rather than that 50 kernels were selected, because the individual differences of kernel size between samples of the same species were still substantial. By weighing, the deviation between sample quantities would be less than it is in this thesis’s measurements. The accuracy of my predictions would have been even more precise if the total sample set would have been bigger than the 500 samples used now.

7.3 Benefits for Evira

The arrival of a handheld near infrared spectroscopy device will open a lot of new possibilities for Evira. At this moment Evira receives samples from ~1500 farms yearly and needs to calculate the national harvest yield based on those samples. If the new device gets sold to the Finnish farmers it becomes possible to receive a lot more data because a lot more farmers would be included in the research.

This would mean, however, that Evira would have to produce new guidelines regarding measurements. The device has GPS and therefore it can combat the trend of farms reducing in numbers but increasing in size. It could be proposed that a farmer would have to do an amount of measurements in his field based on the number of hectares and the location within the field, these coordinates can then easily be checked using satellite data.

GrainSense’s device will be linked to the cloud and this will also help with the calibration of all devices in the form of updates. This means that every farmer will be able to simultaneously produce more accurate results when a better statistical method is found.

7.4 Benefits for the farmers

While the thesis was conducted to answer the question if GrainSense's device would add to the information content of Evira's annual grain survey, it is also good to discuss the benefits for the farmer, since he/she will be the one purchasing and using the device.

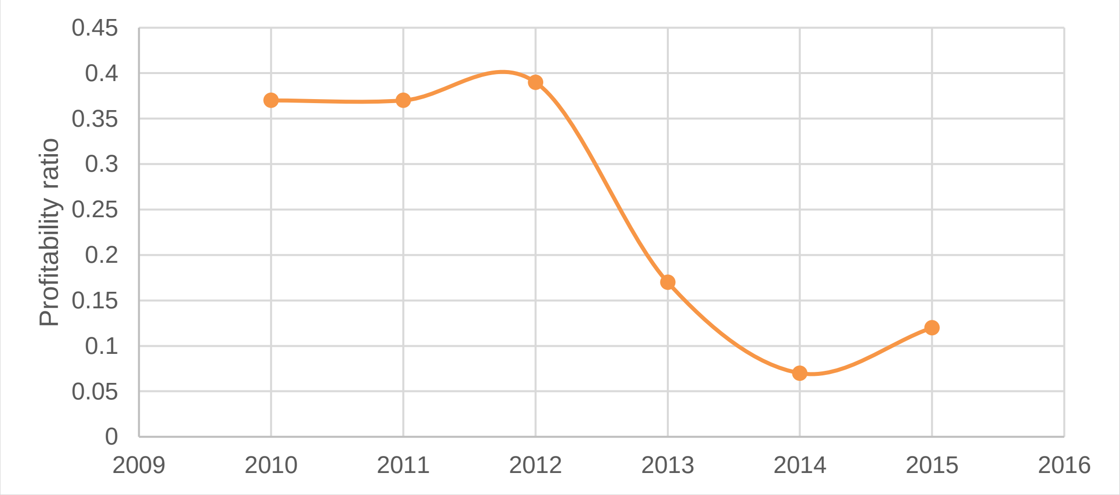

Figure 17. Average profitability ratio of cereal farms in Finland

The profitability ratio is of a farm is calculated by dividing its total income by the sum of costs. When the profitability ratio is 1 all production costs including costs of factors like employees' wages have been covered and the entrepreneur's profit is zero. Consequently, when the profitability ratio is less than 1 it means that the farm is losing money and when its more than 1 that the farm is making a profit. Profitability ratios are often used as a comparison between different years. In Figure 17 you can see that the average farm in Finland is struggling to be profitable and that the profitability ratio dropped drastically since 2012. (Luke, Natural Resources Institute Finland, 2017)

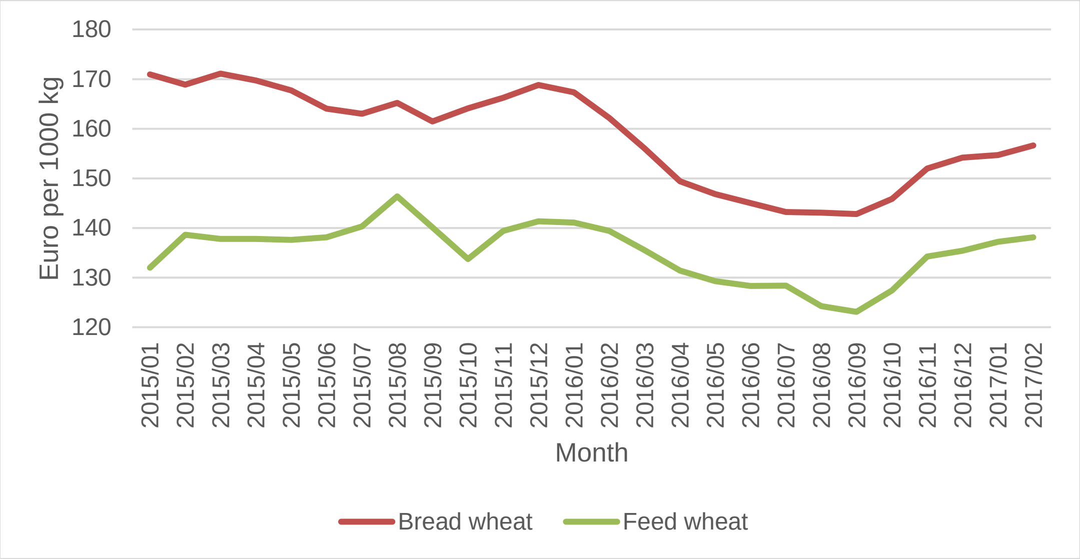

Figure 18. Difference in price per wheat quality

It would benefit the farmer if he/she gets to sort his/her grain more efficiently, in Figure 18 the difference between the price of wheat used for bread and the price of wheat used to feed animals makes this clear. If part of the farmer’s field yields high quality grain and another part yields poor grain because of poor irrigation, this could result in selling the totality of that field’s harvest as feed wheat due to not meeting the standard for bread wheat.

With GrainSense’s device the farmer could measure different places in his/her field and know where to either improve the field conditions, or in the worst case determine which part of his/her field will yield poor quality grain and separate it when harvesting. The frequency of measurements would also improve with this device and that could help optimising operations throughout the lifecycle of the grain. This way the revenue of the farm will increase and this will be a benefit for the profitability of the farm. Precise knowledge over their own crops will also give farmers a stronger position when selling their crops.

Currently when the farmer would like to know about the content of his/her field he/she would have to send a sample to a laboratory to get tested, this is a lengthy process and would take a couple of weeks. The benefit for the farmer of owning his own device would be that measurements could be taken whenever it seems necessary.

When bringing the grain to the grain dryer to prepare the grain for storage a more accurate knowledge of the moisture levels of the grain will save energy which in return is better for the environment

In the long term knowing the yields of their own farm will aid farmers with providing means to prove environmental compliance such as minimised run-off.

7.5 Further research

The results of this thesis taught us a lot about how to continue in the future with this project. My suggestion is to use the device in parallel with Evira one of the upcoming year so that the results of both Evira’s Kjeldahl experiments and their NIR measurements can be directly compared to the performance of the GrainSense device.

A possible next step to get good reference data for the GrainSense device might be to take samples and mix them well so that the samples are completely homogeneous. Then half of those samples will be sent to a certified laboratory and the other half will stay with GrainSense for comparative measurements.

Also, there is more than one way to fit a model, for this thesis it was only possible to implement partial least square regression but for the future I suggest a comparison between different methods such as for example artificial neural networks.

On the long term, it would be good for Evira to test this device with a pilot group of farms with different standard outputs to see if owning this device will make a difference in revenue between them and regular farms.

8 Conclusion

The aim of this thesis was to show how measuring grain using a handheld device could impact the quality of the research data Evira collects annually. While this has not been tested on a large scale the device produced by GrainSense shows promise for the future. The results measured with the device were precise (as shown in Figure 13) but showed a larger deviation from the data Evira obtained with the nitrogen determination of the Kjeldahl method. This was due to sources of error that were already known beforehand (sample size, time difference between measurements and not weighing the sample, so the deviation from the results were already expected to be higher. In a sense this thesis therefore is something to be worked upon, with better samples of both better size and quality.

The test did show a clear difference between what happens if you try to make a model based on 50 samples and a model based on 150. The amount of random error reduces the moment you increase your sample size.

With this in mind it can be concluded that Evira would benefit from farmers owning their own measuring data because if every farmer would own a device like the one Grain-Sense produced the accuracy would improve greatly. Another way how the accuracy would be improved is that a farmer can measure different spots in his field, this means that the data that would be sent per farmer would also improve.

A sub-theme of this thesis was how this device can change the lives of farmers, this theme was important to the thesis because there needs to be a motive for the farmers to purchase their own measurement equipment. in Figure 18 it is shown that the difference in price between bread wheat and feed wheat is between the 20 and 30 euros per 1000 kg depending on the date. With around 4 billion kg of grain being produced in Finland annually it would give farmers an increase of revenue if they can separate their high-quality crops from low quality crops, and with that increase the competitiveness of their farms.

9 References

Jokela, A . (2016). *The number of farms is shrinking but production is not*, Available at: <https://www.luke.fi/en/the-number-of-farms-is-shrinking-but-production-is-not/> [Accessed 03/05/2017].

Batey, I. (2017). The Diversity of Uses for Cereal Grains. In C. Wrigley, I. Batey, & D. Miskelly, *Cereal Grains Assessing and Managing Quality*. Australia: Woodhead Publishing. pp. 41-46.

Blamire, J. (2003). *Kjeldahl Method*., Available at: <http://www.brooklyn.cuny.edu/bc/ahp/SDKC/Chem/SD_KjeldahlMethod.html > [Accessed 03/05/2017]

Anon, *Crop production statistics*. (2017)., Available at: <http://stat.luke.fi/en/crop-production-statistics> [Accessed 03/05/2017]

Eurostat. (2017). *Glossary:Satandard output (SO)*. Available at: <http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Standard_output_(SO) > [Accessed 03/05/2017]

FAO. (2011). *Rural structures in the tropics. Design and development.* Rome: CTA.

Finnish Food Safety Authority Evira. (2016). *Statistics about the quality of the grain crop* Available at: <https://www.evira.fi/en/plants/cultivation-and-production/cereals/statistics-about-the-quality/ > [Accessed 03/05/2017]

Finnish Food Safety Authority Evira. (2017). *Grain quality 2015* . Available at: <https://www.evira.fi/en/plants/cultivation-and-production/cereals/statistics-about-the-quality-of-the-grain-crop/grain-quality-2015/> [Accessed 03/05/2017]

GrainSense Oy. (2016). *Finnish agri-tech start-up, GrainSense, secures EUR 1.4 M in equity financing*. Available at: <http://www.vttresearch.com/media/news/grainsense> [Accessed 03/05/2017]

Greenhouse gas inventory unit. Statistics Finland . (2016). *Finland's greenhouse gas emissions decreased further*. Available at: <http://www.stat.fi/til/khki/2015/khki_2015_2016-05-25_tie_001_en.html> [Accessed 03/05/2017]

International Standards Office. (2006). *ISO 20483:2006 Cereals and pulses. Determination of the nitrogen content and calculation of the crude protein content. Kjeldahl method.* Geneva: ISO.

Koehler, P., and Wieser, H. (2013). Chapter 2 Chemistry of Cereal Grains. In M. Gobbetti, & M. Gänzle, *Handbook on Sourdough Biotechnology*. New York: Springer. pp. 11-45.

Luke, Natural Resources Institute Finland. (2017). *Profitability ratio by Production type*. Available at: <https://portal.mtt.fi/portal/page/portal/economydoctor/farm_economy/timeline/profitability_ratio_by_production_type> [Accessed 03/05/2017]

Metrohm AG. (2013). *A guide to near-infrared spectroscopic analysis of industrial munufacturing processes.* Herisau: Metrohm AG. Available through: <http://www.mep.net.au/wpmep/wp-content/uploads/2013/05/MEP_Monograph_NIRS_81085026EN.pdf > [Accessed 03/05/2017]

Natural Resources Institute Finland. (2016). *Use of Crops on Farms by Data, Species and Harvest year*. Available at: <http://statdb.luke.fi/PXWeb/pxweb/en/LUKE/LUKE__02%20Maatalous__04%20Tuotanto__28%20Maatilojen%20sadonkaytto/01_Maatilojen_sadonkaytto.px/information/informationView/?rxid=c90ab152-f296-4512-8187-6a43ebd7b98e> [Accessed 03/05/2017]

Natural Resources Institute Finland. (2017). *Structure of agricultural and horticultural enterprises* . Available at: <http://statdb.luke.fi/PXWeb/pxweb/en/LUKE/LUKE__02%20Maatalous__02%20Rakenne__02%20Maatalous-%20ja%20puutarhayritysten%20rakenne/07a_Maatalous_ja_puutarhayrit_lkm_tal_koko.px/information/informationView/?rxid=b93bd1a3-b5db-4647-9077-85afce4b619d> [Accessed 03/05/2017]

Partela, A. (2017). *The grain and potatoe harvest will suffice for domestic consumption.* Available at: <https://www.luke.fi/en/news/the-grain-and-potato-harvest-will-suffice-for-domestic-consumption/> [Accessed 03/05/2017]

Rajaniemi, M., Mikkola, H., and Ahokas, J. (2011). Greenhouse gas emissions from oats, barley, wheat and rye production. *Biosystem Engineering*, pp.189-195.

Stuart, B. H. (2004). *Infrared spectroscopy: Fundamentals and Applications.* Sydney: John Wiley & Sons Ltd.

Tobias, R. D. (1995). *Focus areas.* Available at: <https://support.sas.com/rnd/app/stat/papers/abstracts/pls.html> [Accessed 03/05/2017]

Appendices

Appendix 1: R-code for oat

```
setwd("F:/Thesis/R")
#source('http://users.metropolia.fi/~velimt/R/DOE_functions_v5.r')

# read in the spectral data
Abs  <- read.table("absorbance_kaura_V2.txt", header = TRUE)

# detect outliers using principal component analyse
pca <- prcomp(Abs[,4:79])
T   <- pca$x
plot(T[,1],T[,2],pch='', main = "PCA Oat samples")
text(T[,1],T[,2],1:588)

# Take means of the replicates
Abs_mean <- aggregate(Abs[, -c(1:3)], by = list(Abs$HVIO_kaura), mean)
names(Abs_mean)[1] <- 'HVIO'

# Remove the outlier
Abs_mean <- Abs_mean[-38,]

# Read in the protein data and remove the outlier
C <- read.table("Kaura_protein_survey_data.txt", header = TRUE)[-38,]

# Determine the training set and the test set
itrain <- 1:120
itest  <- (1:146)[-itrain]

# Load the pls Package
library(pls)

# choose the dimension for the pls model
# (first time the script is run pick an random number,
# second time it's the global minimum dimension)
Dim = 16

# Create the pls model

Xy <- data.frame(y=C$Protein,X=I(as.matrix(Abs_mean[,2:77])))

plsModel <- plsr(y~X, data=Xy[itrain,],validation='CV')
print(summary(plsModel))

# plot the RMSEP from the model to find the global minimum
plot(RMSEP(plsModel),main="Oat model", xlim = c(0,40))

proteinpred <- predict(plsModel,newdata=Xy[itest,])[,,Dim]

Proteinfit <- predict(plsModel)[,,Dim]
plot(C$Protein[itrain],Proteinfit,
     main = "PLS model oat    [R-squared:  0.9153]",
     xlim = c(8,16),
     ylim = c(8,16),
     xlab = "Measured",
     ylab = "Calculated",
     col='orange',
     pch=16)
```

```
points(C$Protein[itest],proteinpred,col='red',pch=16)
#add a legend
legend("topleft",bty='n',
       inset = .02,
       c("measured","predicted",'y=x'),
       col=c('orange','red','blue'),
       pch=c(16,16,NA),
       lwd=c(NA,NA,2),
       horiz=FALSE)
# y=x line
abline(c(0,1),col='blue', lwd=2)

print(summary(lm(Proteinfit~C$Protein[itrain])))

#plot(Xy[itest,'y'],pred,xlim=c(8,15),ylim=c(8,15))
#points(C$Protein[itest],pred,col='red',pch=16)
#abline(c(0,1))

# determine accuracy of prediction
print(rms(Xy[itest,'y']-proteinpred))
# convert to % deviation
print((rms(Xy[itest,'y']-proteinpred)/mean(Xy$y))*100)

# Create a histogram of the data
hist(C$Protein,
     main="Protein distribution of oat samples",
     xlab="Percent of protein",
     border="blue",
     col="orange",
     breaks=12, #Break equals approximate square root of amount of
samples
     xlim=c(8,16),
     freq = FALSE)# we want to see density rather than frequency

# Read in the wavelength data and plot the spectral data
wl <- t(read.table("Wavelengths.txt", header = FALSE))
matplot(t(wl),t(Abs_mean[,2:77]),type='l',
        main = "Spectral data oat",
        xlab = "Wavelength [nm]",
        ylab = "Absorbance")
```

Appendix 2: R-code for barley

```
setwd("F:/Thesis/R")
#source('http://users.metropolia.fi/~velimt/R/DOE_functions_v5.r')

# read in the spectral data
Abs  <- read.table("absorbance_ohra_V2.txt", header = TRUE)

# detect outliers using principal component analyse
pca <- prcomp(Abs[,4:79])
T   <- pca$x
plot(T[,1],T[,2],pch='',main = "PCA Barley samples")
text(T[,1],T[,2],1:636)

# Take means of the replicates
Abs_mean <- aggregate(Abs[, -c(1:3)], by = list(Abs$HVIO_ohra), mean)
names(Abs_mean)[1] <- 'HVIO'

# Remove the outlier
Abs_mean <- Abs_mean[-21,]

# Read in the protein data and remove the outlier
C <- read.table("Ohra_protein_survey_data.txt", header = TRUE)[-21,]

# Determine the training set and the test set
itrain <- 1:130
itest  <- (1:158)[-itrain]

# Load the pls Package
library(pls)

# choose the dimension for the pls model
# (first time the script is run pick an random number,
# second time it's the global minimum dimension)
Dim = 18

# Create the pls model
Xy <- data.frame(y=C$Protein,X=I(as.matrix(Abs_mean[,2:77])))

plsModel <- plsr(y~X, data=Xy[itrain,],validation='CV')
print(summary(plsModel))
proteinpred <- predict(plsModel,newdata=Xy[itest,])[,,Dim]

Proteinfit <- predict(plsModel)[,,Dim]
plot(C$Protein[itrain],Proteinfit,
     main = "PLS model barley    [R-squared:  0.9027]",
     xlim = c(8,15),
     ylim = c(8,15),
     xlab = "Measured",
     ylab = "Calculated",
     col='orange',
     pch=16)

points(C$Protein[itest],proteinpred,col='red',pch=16)
legend("topleft",bty='n',
       inset = .02,
       c("measured","predicted",'y=x'),
       col=c('orange','red','blue'),
       pch=c(16,16,NA),
       lwd=c(NA,NA,2),
       horiz=FALSE)
abline(c(0,1),col='blue', lwd=2)

print(summary(lm(Proteinfit~C$Protein[itrain])))

points(C$Protein[itest],proteinpred,col='red',pch=16)

# determine accuracy of prediction
print(rms(Xy[itest,'y']-proteinpred))
# convert to % deviation
print((rms(Xy[itest,'y']-proteinpred)/mean(Xy$y))*100)


# Create a histogram of the data
hist(C$Protein,
     main="Protein distribution of barley samples",
     xlab="Percent of protein",
     border="blue",
     col="orange",
     breaks=13,#Break equals approximate square root of amount of sam-
ples
     freq = FALSE)# we want to see density rather than frequency

# plot the RMSEP from the model to find the global minimum
plot(RMSEP(plsModel),main="Barley model", xlim = c(0,40))

# Read in the wavelength data and plot the spectral data
wl <- t(read.table("Wavelengths.txt", header = FALSE))
matplot(t(wl),t(Abs_mean[,2:77]),type='l',
        main = "Spectral data barley",
        xlab = "Wavelength [nm]",
        ylab = "Absorbance")
```

Appendix 3: R-code for rye

```
setwd("F:/Thesis/R")
#source('http://users.metropolia.fi/~velimt/R/DOE_functions_v5.r')

# read in the spectral data
Abs  <- read.table("absorbance_ruis_V2.txt", header = TRUE)

# detect outliers using principal component analyse
pca <- prcomp(Abs[,4:79])
T   <- pca$x
plot(T[,1],T[,2],pch='',main = "PCA Rye samples")
text(T[,1],T[,2],1:196)

# Take means of the replicates
Abs_mean <- aggregate(Abs[, -c(1:3)], by = list(Abs$HVIO_ruis), mean)
names(Abs_mean)[1] <- 'HVIO'

# Remove the outlier
Abs_mean <- Abs_mean[-44,]

# Read in the protein data and remove the outlier
C <- read.table("Ruis_protein_survey_data.txt", header = TRUE)[-44,]

# Determine the training set and the test set
itrain <- 1:40
itest  <- (1:48)[-itrain]

# Load the pls Package
library(pls)

# choose the dimension for the pls model
# (first time the script is run pick an random number,
# second time it's the global minimum dimension)
Dim = 11

# Create the pls model
Xy <- data.frame(y=C$Protein,X=I(as.matrix(Abs_mean[,2:77])))

plsModel <- plsr(y~X, data=Xy[itrain,],validation='CV')
print(summary(plsModel))
proteinpred <- predict(plsModel,newdata=Xy[itest,])[,,Dim]

Proteinfit <- predict(plsModel)[,,Dim]
plot(C$Protein[itrain],Proteinfit,
     main = "PLS model rye    [R-squared:  0.7469]",
     xlim = c(8,16),
     ylim = c(8,16),
     xlab = "Measured",
     ylab = "Calculated",
     col='orange',
     pch=16)
```

```r
points(C$Protein[itest],proteinpred,col='red',pch=16)
legend("topleft",bty='n',
       inset = .02,
       c("measured","predicted",'y=x'),
       col=c('orange','red','blue'),
       pch=c(16,16,NA),
       lwd=c(NA,NA,2),
       horiz=FALSE)
abline(c(0,1),col='blue', lwd=2)

print(summary(lm(Proteinfit~C$Protein[itrain])))


points(C$Protein[itest],proteinpred,col='red',pch=16)

# determine accuracy of prediction
print(rms(Xy[itest,'y']-proteinpred))
# convert to % deviation
print((rms(Xy[itest,'y']-proteinpred)/mean(Xy$y))*100)

# Create a histogram of the data
hist(C$Protein,
     main="Protein distribution of rye samples",
     xlab="Percent of protein",
     border="blue",
     col="orange",
     breaks=7,#Break equals approximate square root of amount of sam-
ples
     freq = FALSE)# we want to see density rather than frequency

# plot the RMSEP from the model to find the global minimum
plot(RMSEP(plsModel),main="Rye model", xlim = c(0,30))

# Read in the wavelength data and plot the spectral data
wl <- t(read.table("Wavelengths.txt", header = FALSE))
matplot(t(wl),t(Abs_mean[,2:77]),type='l',
        main = "Spectral data rye",
        xlab = "Wavelength [nm]",
        ylab = "Absorbance")
```

Appendix 4: R-code for wheat

```
setwd("F:/Thesis/R")
#source('http://users.metropolia.fi/~velimt/R/DOE_functions_v5.r')

# read in the spectral data
Abs  <- read.table("absorbance_vehna_V2.txt", header = TRUE)

# detect outliers using principal component analyse
pca <- prcomp(Abs[,4:79])
T   <- pca$x
plot(T[,1],T[,2],pch='',main = "PCA Wheat samples")
text(T[,1],T[,2],1:584)

# Take means of the replicates
Abs_mean <- aggregate(Abs[, -c(1:3)], by = list(Abs$HVIO_vehna), mean)
names(Abs_mean)[1] <- 'HVIO'
Abs_mean <- Abs_mean

# Read in the protein data
C <- read.table("Vehna_protein_survey_data.txt", header = TRUE)

# Determine the training set and the test set
itrain <- 1:120
itest  <- (1:146)[-itrain]

# Load the pls Package
library(pls)

# choose the dimension for the pls model
# (first time the script is run pick an random number,
# second time it's the global minimum dimension)
Dim = 21

# Create the pls model
Xy <- data.frame(y=C$Protein,X=I(as.matrix(Abs_mean[,2:77])))

plsModel <- plsr(y~X, data=Xy[itrain,],validation='CV')
print(summary(plsModel))
proteinpred <- predict(plsModel,newdata=Xy[itest,])[,,Dim]

Proteinfit <- predict(plsModel)[,,Dim]
plot(C$Protein[itrain],Proteinfit,
     main = "PLS model wheat    [R-squared:  0.9755]",
     xlim = c(8,16),
     ylim = c(8,16),
     xlab = "Measured",
     ylab = "Calculated",
     col='orange',
     pch=16)

points(C$Protein[itest],proteinpred,col='red',pch=16)
legend("topleft",bty='n',
       inset = .02,
       c("measured","predicted",'y=x'),
       col=c('orange','red','blue'),
       pch=c(16,16,NA),
       lwd=c(NA,NA,2),
       horiz=FALSE)
```

```
abline(c(0,1),col='blue', lwd=2)


print(summary(lm(Proteinfit~C$Protein[itrain])))


points(C$Protein[itest],proteinpred,col='red',pch=16)

# determine accuracy of prediction
print(rms(Xy[itest,'y']-proteinpred))
# convert to % deviation
print((rms(Xy[itest,'y']-proteinpred)/mean(Xy$y))*100)

# Create a histogram of the data
hist(C$Protein,
     main="Protein distribution of wheat samples",
     xlab="Percent of protein",
     border="blue",
     col="orange",
     breaks = 13,#Break equals approximate square root of amount of
samples
     xlim=c(8,18),
     freq = FALSE)# we want to see density rather than frequency

# plot the RMSEP from the model to find the global minimum
plot(RMSEP(plsModel),main="Wheat model", xlim = c(0,40))

# Read in the wavelength data and plot the spectral data
wl <- t(read.table("Wavelengths.txt", header = FALSE))
matplot(t(wl),t(Abs_mean[,2:77]),type='l',
        main = "Spectral data wheat",
        xlab = "Wavelength [nm]",
        ylab = "Absorbance")
```