Effect of Milling on Electrostatic Separation and Modeling Protein and Starch Content of Flour Fractions

by

Michael Vitelli

A thesis

presented to the University of Waterloo

in the fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Chemical Engineering

Waterloo, Ontario, Canada, 2017

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The objective of this research is to establish the effects of different milling techniques on the solvent-free electrostatic separation process for navy bean flour as well as to develop a model based on near infrared and fluorescence data to determine protein and starch content of the protein- and starch-enriched fractions using multivariate methods (i.e. partial least squares regression). Data fusion was used to combine the NIR and fluorescence spectra to try to achieve a model that had better predictability for protein and starch content.

Protein content was measured using Kjeldahl digestion and starch content was measured using a dinitrosalicylic (DNS) acid array. The samples used in the NIR model are navy bean flour fractions from the electrostatic separation and the raw navy bean flour. There are 102 samples that are split in calibration (82 samples) and validation (20 samples) sets. The protein-enriched samples are collected from the electrostatic plate while the starch-enriched fractions are collected from the bottom of the electrostatic separator. The acquisition of reproducible infrared and fluorescence data from powder samples was successfully achieved.

The pin milled navy bean flour had an average particle size almost three times smaller than the regular milled navy bean flour which could have contributed to the a high protein content (40.7%) of the protein-enriched fraction. The regular milled flour had a much higher protein extraction under optimum conditions but could only achieved a lesser protein content (32.5%) for the protein-enriched fraction. The regular milled navy bean flour also seemed to have particles disaggregate in the triboelectric charging process.

Multivariate methods and pre-treatment techniques were compared for the NIR spectra of the navy bean flour fractions from electrostatic separation to measure the protein and starch content.

The best method used Multiplicative Scatter Correction (MSC) pre-treatment with PLS regressions and had R^2 values of prediction of 0.965 and 0.912 for protein and starch content, respectively.

The N-way partial least squares (NPLS) regression was still a good model seeing as the R^2 values of prediction for starch and protein content were 0.946 and 0.885, respectively. Two fluorophores were observed in navy bean flour: tryptophan and an unknown peak. It was observed that the starch model using the fluorescence dataset was highly correlated to the model's predicted protein content (R^2 of 0.978). The protein content model was better calibrated using the training set as well as providing a better prediction using the validation set for both NIR and fluorescence spectra.

Data fusion was achieved by combining the NIR and unfolded fluorescence spectra of the navy bean flour fractions. The individual techniques had undergone pre-treatment separately and yielded the best model for determining protein content. Starch content was best determined using only the NIR spectra.

Acknowledgements

First and foremost, I would like to thank Dr. Raymond Legge for giving me the opportunity to work in his lab and the support he has shown me over the course of this work. I am also grateful for the financial support he gave me during my second year as a grad student provided through his NSERC operating grant.

For his assistance in finding and implanting the methodology for finding starch content in flour samples and his countless hours of running experiments, I would like to thank Andrew Assatory.

For his assistance in helping me with the front-faced fluorescence set-up, I would like to thank Joseph Ferr.

For guidance and encouragement over the course of this work, I would like to thank Amin Rajabzadeh.

For endless hours of conversation, and the unspoken rule for it never to be about work, I would like to thank Rob Claridge.

I would like to thank my parents, Joe and Cindy, and my siblings, Steven and Alyssa, for their support throughout the work.

Finally, I would like to thank Dr. Hector Budman and Prof. Bill Anderson who graciously accepted to be my thesis readers. Their insight, comments, and corrections are greatly appreciated.

Table of Contents

List of Figures

List of Tables

List of Nomenclature

1D	Savitzky-Golay first derivative
2D	Savitzky-Golay second derivative
A	Sample absorbance
ABS	Acrylonitrile butadiene styrene
ANN	Artificial neural network
B	Blank absorbance
BPNN	Back propagation neural network
C	Capacitance
DA	Discriminant analysis
DAD	Diode array detector
DBD	Dielectric barrier discharge
ε_o	Absolute gas permittivity
EEM	Excitation-emission matrix
EPLS	Ensemble partial least squares
IgG	Immunoglobin G
IVCD	Integrate viable cell density
k	Standard calibration curve slope
k_c	Charging efficiency
L_o	Characteristic length of particle charging
LSNV	Localized standard normal variate
MBPLS	Multi-block partial least squares
MLR	Multiple linear regression
MSC	Multiplicative scatter correction
n	Number of contacts
NIR	Near infrared
NMR	Nuclear magnetic resonance
NPLS	n-way partial least squares
OSC	Orthogonal scatter correction

PARAFAC	Parallel factor analysis
PCA	Principal component analysis
PCR	Principal component regression
PET	Polyethylene terephthalate
PLS	Partial least squares
PVC	Polyvinyl chloride
PP	Polypropylene
PS	Polystyrene
PVP	Polyvinyl propylene
Δq	Chang in charge on the particle
q	Particle charge before contact
RMSEC	Root mean squared error of calibration
RMSECV	Root mean squared error of cross validation
RMSEV	Root mean squared error of validation
S	Contact area
SNV	Standard normal variate
SVM	Smart vector machine
UV	Ultraviolet
V	Total potential difference / volume
V_b	Potential difference arising from surrounding particles
V_c	Potential difference from work functions
V_e	Potential difference from image charge
V_{ex}	Potential difference from other electric fields
VIP	Variable importance in the projection
w	Mass of sample
z_o	Critical gap

Chapter 1

Introduction

A novel, solvent-free triboelectric separation method has been developed for the production of protein- and starch-enriched fractions for agricultural flours. This technique separates the triboelectrically charged flour particulates using an electrostatic field from a high voltage plate. The advantages of using this separation method include that the protein is recovered in its native state and there is no downstream requirement for solvent removal, which is required in other approaches.

Triboelectric charging occurs when two materials are brought into contact with each other and charge is gained for one material due to friction. The amount of charge gained on a particulate from triboelectric charging depends on factors such as surface conditions, area of contact, speed of rubbing, the materials involved, and humidity (Matsusaka *et. al* 2010). Electrostatic separation involves the separation of charged particles under the influence of an electric field. The electric field is created by one or more high voltage electrodes and the particles usually enter the separator after being charged from triboelectric or corona processes.

It is possible to gain insight on the chemical composition and properties using spectroscopic techniques. Near infrared spectroscopy is a fast, non-invasive technique that requires little to no sample pre-treatment or preparation. It can be used to determine the chemical composition of components in a variety of complex organic samples through vibrational energy resulting from molecular bonds absorbing the infrared light (Workman & Weyer 2012). Fluorescence spectroscopy is also a non-destructive analysis that can accurately measure component concentrations at parts per billion (Christensen *et al.* 2006). Molecules that can act as fluorophores, whose chemical structure allows it to be excited by an external photon and then emit the photon while relaxing back to its ground state, are found in fluorescence spectra. The molecules that fluoresce mainly involve aromatic structures but can also be molecules with carbonyl groups or highly conjugated double bonds (Patnaik 2004).

The objective of this research was to establish the effects of different milling techniques on the solvent-free electrostatic separation process for navy bean flour as well as to develop a model based on near infrared and fluorescence data to determine protein and starch content of the protein- and starch-enriched fractions using multivariate methods (i.e. partial least squares regression). Acquisition of reproducible infrared and fluorescence data from powder samples was also an objective. After the models for the near infrared and fluorescence spectra were developed, they were compared to a model that used a data fusion approach. The advantages of data fusion are improved detection, an increase in accuracy and reliability, and extending the available data (Khaleghi *et al.* 2013).

Chapter 2

Literature Review

2.1 Triboelectric Charging

Triboelectric charging occurs when two different materials are brought into contact where electric charge is transferred from one material to the other. The materials become electrically charged due to friction. The magnitude of the charge is affected by surface conditions, area of contact, speed of rubbing, the materials involved, and humidity (Matsusaka *et. al* 2010).

The charge of materials that are brought into contact with each other is related to the work function of each material. Work function is defined as the energy required to remove an electron from the surface of a material. Work function of materials is unique since it depends on the electronic energy levels of the particular material due to its chemical composition. Smooth surfaces with a great amount of contact pressure and relative motion result in greater changes in charge.

Whether the material retains the charge depends on its conductivity and the availability for the charge to be grounded. Grounding the material allows electrons to flow from or to the material returning it to its normal state.

Charged particles are an instrumental part of various industrial applications. Applications that use triboelectric charging equipment include electrostatic separation (Zenkiewicz *et al.* 2015), powder coating (Dastoori *et al.* 2005), and electrophotography (Schein *et al.* 1999). Therefore, understanding the characteristics of charging the particles is important to the powder handling process and optimization of parameters to improve the application's performance.

2.1.1 Triboelectric Series

A triboelectric series is a list of materials arranged according to their charge intensity. The materials are ordered from negatively charged (electron accepters) to positively charged (electron donors). Triboelectric series can be used to determine the material to gain the greatest charge or electrostatic separation processes. Figure 2-1 (Licari 1988) is an example of a triboelectric series. Teflon was chosen to be the material of the triboelectric charger (method and design in section 3.3) because it most readily donates electrons as it is on the bottom of the triboelectric series.

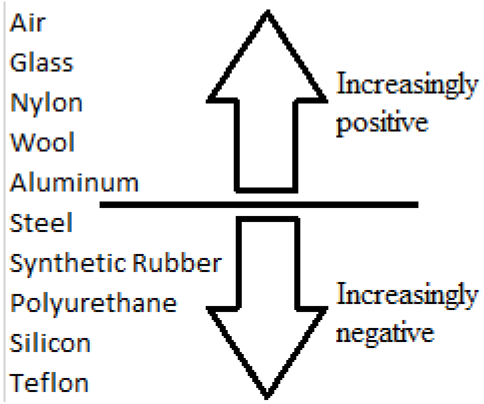


Figure 2-1: Triboelectric series of common materials used in triboelectric charging (adapted from Licari (1988)).

Hyun *et al.* (2008) created a tribocharger which combined a vertical reciprocator and various charger materials. Their tribocharger was designed to charge plastics by a transfer from rotating motion to reciprocation of a cam axis into a charging bottle. The charging properties of plastics were measured and a triboelectric series was obtained. The tribocharger was used to predict material separation for the recycling of waste plastic.

Zenkiewicz *et al.* (2015) developed a tribocharger series for three biodegradable polymers: polylactide (PLA), polycaprolactone (PCL), and poly(3-hydroxybuterate-co-hydroxyvalerate) (PHBV). Two triboelectric chargers were used in the study. A mechanical tribocharger was used and employed the mutual friction between particles interacting with each other and through the particle-tribocharger wall contacts. A fluidized bed was also used to triboelectrically charge particles as the movement causes friction. It was found that PLA/PCL mixtures can be electrostatically separated. These two biodegradable biopolymers have the ability to be multi-processed without considerable deterioration.

2.1.2 Triboelectric Charge Modelling

Mizutani *et al.* (2013) developed an excellent derivation of a triboelectric charging model. The following summarizes the steps taken to reach an equation relating the length of the triboelectric charger with particle charge.

The contact region between a particle and the triboelectric charger wall can be considered the same as a capacitor. Therefore, the charge induced by the impact can be represented by (1).

$$\Delta q = k_c C V \quad (1)$$

where: Δq: change in charge of the particle

k_c: charging efficiency

C: capacitance

V: total potential difference

The capacitance (2) relates to the contact area of the particle and the critical gap during the particle-wall impact as shown in Figure 2-2.

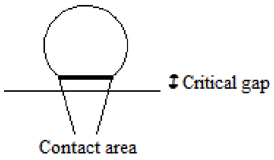


Figure 2-2: Particle wall impact variables adapted from Mizutani (2013).

$$C = \frac{\varepsilon_o S}{z_o} \quad (2)$$

where: ε_o: absolute permittivity of gas

S: contact area

z_o: critical gap including geometrical factors between contact bodies

The total potential difference, shown in equation (3), relates to the particle wall contact and the potential difference arising from other particles and external electric fields.

$$V = V_c - V_e - V_b + V_{ex} \quad (3)$$

where: V_c: potential difference obtained based on work functions

$V_e = k_e q$: potential difference arising from image charge

q: particle charge before contact

V_b: potential difference arising from the space of charge induced by the surrounding particles, negligible under dilute conditions

V_{ex}: potential difference arising from other electric fields

The amount of charge that is induced per impact decreases as the number of impacts increase. Therefore, the particle charge is generated through repeated impacts.

$$\frac{dq}{dn} = k_c C V \quad (4)$$

where: n: number of contacts

Using equations (1-4) under dilute conditions (low concentration of particles in the carrier gas), the equation for charge transfer becomes the following:

$$\frac{dq}{dn} = -\frac{k_c k_e \varepsilon_o S}{z_o} q + \frac{k_c \varepsilon_o S (V_c + V_{ex})}{z_o} \quad (5)$$

Assuming the initial conditions of $q=q_o$ when $n=0$ and it is possible to solve the differential equation.

$$q(n) = q_o e^{\frac{-n}{n_o}} + q_\infty \left(1 - e^{\frac{-n}{n_o}}\right) \quad (6)$$

where: $n_o = \frac{z_o}{k_c k_e \varepsilon_o S}$

$q_\infty = \frac{V_c}{k_e} + \frac{z_o}{k_e} V_{ex}$

When the particles are travelling on a plate, it can be assumed the frequency of the particle wall impacts per unit plate length is constant. Therefore, the length of the plate also has a relationship with the charge.

$$q(L) = q_o e^{\frac{-L}{L_o}} + q_\infty \left(1 - e^{\frac{-L}{L_o}}\right) \quad (7)$$

where: L_o: characteristic length of particle charging

2.1.3 Further Models of Particle Charge

One assumption from Mizutani *et al.* (2013) work is that the particles have a uniform charge; however, this is not the case as the particle's charge will be localized in the area of contact. Grosshans *et al.* (2017) developed a model that accurately accounts for the charge distribution on the particle during particle-wall and particle-particle collisions implemented in Computational Fluid Dynamics (CFD). It was shown that the behaviour of non-uniform particles was different quantitatively and qualitatively from uniformly charged particles.

Grosshans *et al.* (2016) evaluated the air flow and pipe diameter as parameters affecting the charge on particles flowing through a pipe. Large Eddy Simulations were performed in conjunction with Design of Experiments (DoE) methodology. It was determined that decreasing the flow rate greatly reduces the charge on the particles. This is because as the flowrate is reduced the number of contacts will decrease as the particles enter laminar flow. The diameter of the pipe had a lesser effect. However, as the pipe diameter increases, the charge on the particles does decrease.

2.2 Electrostatic Separation

Electrostatic separation is a dry, non-destructive technique used mainly in the plastic and food industries. In the plastic industry, electrostatic separation is mainly used for recycling pure plastic components from a milled mixture of components (Inculet *et al.* 2017; Nadjem *et al.* 2017; Zeghloul *et al.* 2016) . In the food industry, electrostatic separation is used to enrich a flour's components such as protein and fibre (Jafari *et al.* 2016; Tabtabaei *et al.* 2016a; Tabtabaei *et al.* 2016b; Tabtabaei *et al.* 2017).

Electrostatic separation involves the separation of charged particles under the influence of an electric field. The electric field is created by one or more high voltage electrodes and can be charged either positively or negatively depending on the components to be separated. The particles gain a charge before entering the separator usually through triboelectric or corona processes.

There are two main types of electrostatic separators. The first uses free-falling particles that come into contact with the electrostatic field. Besides the electrostatic force from the electrode, the forces of gravity and the friction between the charged particles and the carrier gas play an important role in the separation process. Therefore, the height of the separator chamber, distance from the particle input to the electrode, the particle sizing, and the electrode voltage have the greatest effect on the separation of the charged particles. The other type of separator involves a carrier gas passing the particles horizontally where they come into contact with the electrostatic field from the electrode. Therefore, the speed of the carrier gas is an additional variable affecting separation.

2.2.1 Electrostatic Separation of Plastics

Inculet *et al.* (2017**)** used electrostatic separation to separate polyvinyl chloride (PVC) and polyethylene terephthalate (PET) mixtures for recycling the pure components. A tower was designed so that the charged particles would be in free-fall and come into contact with an electrostatic field coming from two high voltage electrodes. One electrode had a positive charge while the other's was negative. The collection bins were at the bottom of the tower and the particulates end up in a particular bin depending on the deflection resulting from the electrostatic field. It was concluded that from a 50/50 mixture of PVC and PET, it was possible to recover 92.7% and 92.9 % of each component respectively.

Nadjem *et al.* (2017) separated polyvinyl propylene (PVP) and polypropylene (PP) granular materials treated with dielectric barrier discharge (DBD) prior to triboelectric charging. DBD modifies the triboelectric properties of the particles and was carried out in a plasma reactor consisting of two aluminum plate electrodes as shown in Figure 2-3. In the free-fall electrostatic separator the electrodes were composed of aluminum and set to voltages of +/- 30 kV. The best enhancements of triboelectric properties involved using DBD for short periods of time (3 seconds). Under optimal conditions the quantities of PVC and PP obtained were increased by 104% and 30% respectively.

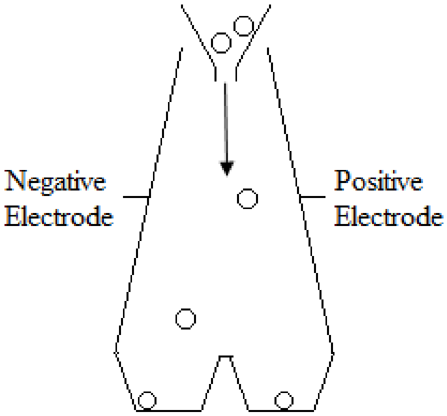


Figure 2-3: Schematic of free-fall electrostatic separator (adapted from Nadjem *et al.* (2017)**)**.

Zeghloul *et al.* (2016) evaluated the effect of particle size on the selective sorting of fine particles using acrylonitrile butadiene styrene (ABS) and polystyrene (PS) mixtures. The goal was to find the optimum design for an industrial electrostatic separator. The parameters that were tested included carrier gas flowrate, electrostatic plate voltage, ad particle size. The particles were charged in a fluidized bed via particle-particle and particle-wall interactions. The electrostatic field was created using rotating disk electrodes. In conclusion, they found that the separation efficiency was best for finely ground particulates (diameter < 1 mm).

2.2.2 Electrostatic Separation in Food

Wang *et al.* (2016) used a variety of dry fractionation techniques for dietary fibre enrichment from pin milled defatted rice. The yields were compared using a two stage electrostatic separation, sieving, and a combination of electrostatic separation and sieving. The particles acquired a charge due to particle-particle and particle-wall collisions occurring with a squared aluminum charging tube. Electrostatic separation took place in a separation chamber with a high voltage aluminum electrode on one side as shown in Figure 2-4. The yield using all three dry fractionation technique was similar, between 20-21%, while recovering 42-48% of the fibre from the original flour.

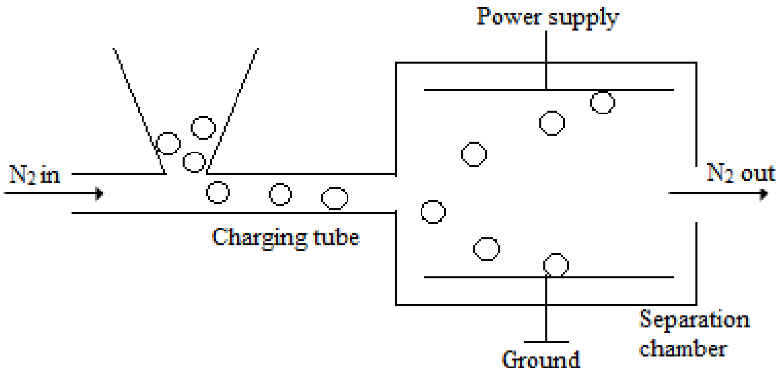


Figure 2-4: Schematic of electrostatic separator (adapted from Wang *et al.* (2015)).

Hemery *et al.* (2011) used electrostatic separation to divide wheat bran into fractions containing high purities of pericarp, testa and aleurone to be used as food ingredients. Coarse bran was obtained using a roller milling process on wheat grains. The coarse bran was then impact milled to become a finer powder. Three stages of free-fall electrostatic separation was used with oppositely charged high voltage electrodes on either side of the separator chamber. The purified fractions resulted in 34.1 and 12.6 % of the starting mass of the pure components.

Schutyser *et al.* (2015) compared dry and wet fractionation techniques when separating legume flour into protein-enriched fractions. Dry fractionation methods, such as electrostatic separation, resulted in lower purity but had a much higher yield. In conclusion, combining dry and wet fractionalization methods increased the purity of the enriched fractions with a smaller decrease in the overall yield from the starting material.

2.2.3 Modeling Electrostatic Separators

Chun *et al.* (2016) modelled a corona charging type electrostatic separator using Artificial Neural Network (ANN) architecture. The schematic of the electrostatic separation process is shown below in Figure 2-5. Electrostatic separation is this approach depends on the centrifugal force due to rotation, lifting force due to the electrode attraction, gravitational force, and pinning force due to electrode attraction from the corona electrode.

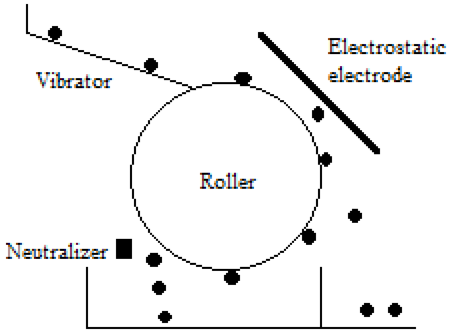


Figure 2-5: Corona charging electrostatic separator (adapted from Masui (1982)).

ANN is a numerical estimation method that simulates the learning and memorizing operation of the brain. It has three types of layers (input, hidden, and output) to compute the complex interactions between neurons.

The input variables used in the experiment were the DC voltage level, the rotation speed of the roller, temperature, the angle of the electrode, the distance between the roller and corona charger, and the distance between the roller surface and electrode. The output of the ANN model was the middling mass product (middling = product that fell between collection bins lowering efficiency) which was compared to the target mass from experiments. The low values of error between the predicted and target masses showed that ANN is a potential tool for modelling the nonlinear electrostatic separation processes.

Labair *et al.* (2017) used COMSOL to model the trajectory of millimeter sized charged particles in a free-fall electrostatic separator. The particles were subjected to electric (from the high voltage electrode) and gravitational forces. The simulation was done on an ABS/PVC mixture containing an average of 4 mm sized granules. The factors found to affect the separation efficiency were determined to be the electrode voltage and the friction between the granules and air. Also, the trajectory of weakly charged granules can be improved by either increasing the voltage of the electrode or increasing the height of the separation chamber.

2.3 Near Infrared Spectroscopy

Near infrared (NIR) spectroscopy is a fast, non-invasive technique that requires little to no sample pre-treatment or preparation. It can be used to determine the chemical composition of components in a variety of complex organic samples. Infrared spectroscopy uses electromagnetic radiation where bonds in a molecule can absorb, transmit, reflect, or scatter infrared light (Workman & Weyer 2012). Molecular spectra result from vibrational energies caused by the absorption of infrared light. The NIR region consists of wavelengths between 780 and 2500 nm (Jerez *et al.* 2010). In this region, functional groups that include hydrogen absorb the wavelengths to create the spectra. Chemical bonds vary in strength and therefore the amount of energy it takes for bonds to vibrate. The different energies correspond to particular wavelengths where peaks in the spectra can be observed.

NIR spectra also commonly undergo pre-treatment effects for baseline correction and curve smoothing. These techniques are mainly used to improve the quantification or classification of the spectra (Berg & Engelsen 2009). Some common examples of pre-treatment are standard normal variate, multiplicative scatter correction, and Savitzky-Golay differentiation. These methods are outlined and discussed in Chapter 5.

The amplitude and width of the peak on the spectra can be calibrated to yield the concentrations of components in a sample. However, most biological samples contain many components which could have overlapping bands rendering it difficult to determine concentrations based on only peak size (Ranzan *et al.* 2014). Therefore, techniques using multivariate analysis have been developed to create models to quantify peaks for the components of the spectra based on a calibration set. The model can then be used to predict the same components of a test or validation set. Commonly used multivariate methods including principal component regression and partial least squares is discussed in Chapter 5.

A review of methods used to classify and quantify components, primarily biological, from NIR spectra can be found in Table 2-1.

Table 2-1: Review of multivariate methods to quantify or classify components in NIR spectra primarily for biological samples.

Sample Type	Measured Component	Pre-treatment	Multivariate Method	Results	Reference
Bio-oil from pine lumber feedstock	Water content	Raw, MSC	PCR, PLS	• Can successfully predict water content between 16 and 36 % with an error less than 2 % • MSC pre-treated data yielded better estimates than the raw data for both PLS and PCR • The best model used MSC pre-treatment with PLS regression: R^2(cal) of 0.971 and R^2(val) of 0.963 • All models had an $R^2 > 0.85$	(Tripathi *et al.* 2009)
Cultivated and wild soybean	Crude protein, crude fat, neutral detergent fibre (NDF), acid detergent fibre (ADF)	MSC	PLS	• Crude fat and crude protein equations are acceptable for quantitative prediction for soybeans (R^2 above 0.9) • NDF and ADF equations are only useful for screening purposes (R^2 approximately 0.75 for both)	(Asekova *et al.* 2016)
Meat Tablet (drug) Wheat	Moisture, fat, and protein content Unspecified Carbon, nitrogen, sulfur content	Raw, 1D, 2D, SNV, MSC, localized SNV (LSNV)	PLS	• Localized SNV uses correction parameters that are estimated over spectral areas • Comparison between a variety of NIR pre-treatment methods for three different datasets • LSNV pre-treatment provided the best overall model for all three datasets • Localized spectral pre-treatments have advantages over full range spectral pre-treatments	(Bi *et al.* 2016)

Warfarin sodium	Crystalline and amorphous fractions	MSC + 1D	PCR, PLS	• Goal: minimize amount of product variation • Linear relationship between changes in the physical form of warfarin sodium and the NIR spectra • Model was very close to actual values obtaining an acceptable method for quantification • Amorphous $R^2 > 0.99$, crystalline $R^2 > 0.98$	(Korang-yeboah *et al.* 2016)
Minced beef adulterated with horsemeat	Horsemeat content	Raw, SNV, MSC, 1D. 2D	PLS	• Utilized the visible wavelength range (400-1000 nm) • Successfully used visible IR to detect and quantify the level of adulteration in horsemeat • All pre-treatment methods had extremely strong correlations ($R^2 > 0.95$) • Best model used the raw data with no pre-treatment	(Kamruzzaman & Makino 2015)
Skim milk powder, non-fat dry milk	Classification Moisture, fat, and protein content	SNV	ANOVA-PCA, Pooled ANOVA, Pooled ANOVA-PLS PLS	• ANOVA-PCA was able to successfully separate the samples by day of analysis, production site, processing temperature, and individual samples into clear categories • The area of protein peaks had the greatest amount of variance between the samples • Calibration models for moisture, fat and protein were not precise, with R^2 of 0.32, 0.6, and 0.78 respectively	(Harnly *et al.* 2014)

Bibasic calcium phosphate – 8 particle sizes (53-300 μm)	Median particle size and logarithmically transformed number of taps	SNV	Multiple linear regression (MLR)	• Quantification using NIR spectra of dry powder • SNV pre-treatment conserves the particle size effect • Tested full spectra and reduced models (wavelength selection) and the full spectra performed better • High correlations for median particle size with full and reduced spectra (R^{2} > 0.95)	(Ely *et al.* 2008)
Soil samples from Belgium and France	Organic carbon, potassium, sodium, magnesium, phosphorus	Maximum normalisation with 1D	PCR, PLS, Back propagation neural network (BPNN)	• Comparison of a variety of multivariate techniques for the best model to determine soil properties from IR/NIR spectra • BPNN outperformed both PLS and PCR for all soil properties • Excellent models for organic carbon and magnesium while the other properties still had good models	(Mouazen *et al.* 2010)
Commercial red wines	Polyphenolic compounds such as malvin, catechin, and quercetin	Normalisation, SNV, MSC, 1D, 2D, combinations	PCR, PLS	• NIR can be used to predict individual polyphenolic compounds • PLS model outperformed PCR • Different pre-treatments were better for different compounds; there was not a single pre-treatment technique that resulted in the best model for all or even most of the compounds	(Vázquez 2014)
Arabica coffee beans	Classification	MSC + 2D combination	PLS-discriminant analysis (DA)	• Goal: detect defects or adulteration in coffee • PLS-DA model was able to discriminate coffee beans geographically and genotypically • Model had classified 94.4 % of the samples correctly • Region model had better separation than genotype model	(Marquetti *et al.* 2016)

Polyester resin from dicarboxylic acids	Acid value and hydroxyl number	Raw	PCR, PLS, BPNN	• Comparing a variety of multivariate methods for the optimal model • BPNN was determined to be the best model for both measured components • BPNN had a faster decrease in predictability as the number of samples decreased • BPNN was successfully employed for monitoring the polyesterification of dicarboxylic acids	(Marengo *et al.* 2004)
Rapeseed biodiesel fuel	Density, viscosity, water content, and methanol content	Mean centering, SNV, MSC, 1D, 2D, combinations, range scattering (for ANN)	PCR, PLS, Artificial neural network (ANN)	• Comparing a variety of multivariate methods as well as pre-treatment methods • ANN was determined to be the superior approach as bio-diesel is a non-linear object • The best NIR pre-treatment method with the dataset was the first or second derivative followed by mean centering	(Balabin *et al.* 2015)
Tablets provided by Roche Pharmaceuticals	Active substance content: bromazepam (Tablet A), clonazepam (Tablet B)	Raw, SNV, MSC, D2, orthogonal scatter correction (OSC), combinations	PCR, PLS	• Comparing a variety of pre-treatment methods using PLS and PCR • All models were highly correlated to the active substance content of both tablets ($R^2 > 0.97$) • PLS provided better predictions than PCR models • Tablet A best pre-treatment: SNV • Tablet B best pre-treatment: SNV and D2	(Chalus *et al.* 2005)

Freeze-dried mannitol-sucrose mixtures	Water content	SNV	PLS	- Created a calibration model using mannitol-sucrose samples which were used to predict water content of other samples (included protein or excipient) - There was a highly linear correlation between the NIR predicted and measured water content	(Grohganz *et al.* 2010)
Soy bean oil	Classification	Raw, offset correction, 1D	PCA, PLS-DA	- Classify expired and non-expired samples of soy bean oil - Acidity and peroxide levels are important indexes of oils that can be determined using NIR - 98% of the samples were successfully classified using PLS-DA indicating the model can be used to evaluate the degree of oxidation in soy bean oil	(Bezerra *et al.* 2016)

0

2.4 Fluorescence Spectroscopy

Fluorescence spectroscopy is a non-destructive analysis that can accurately measure component concentrations at parts per billion (Christensen *et al.* 2006). A fluorophore is a molecule whose chemical structure allows it to be excited by an external photon and then emit the photon while relaxing back to its ground state. The photon is always emitted at a longer wavelength than the wavelength it was excited at. The molecules that fluoresce mainly involve aromatic structures but can also be molecules with carbonyl groups or highly conjugated double bonds (Patnaik 2004). All fluorophores have independent and specific wavelengths for excitation and emission which also result in different peak shapes. Using several emission spectra taken at different excitation wavelengths, an excitation-emission matrix (EEM) can be obtained.

Fluorescence data has two types of scattering which are elastic (Rayleigh) and inelastic (Raman) (Andersen 2005). Rayleigh scattering occurs due to molecules oscillating at a multiple of the incident light frequency. First order Rayleigh scatter would have the same frequency as the incident light and second order would have double the wavelength. Raman scattering occurs since the emitted light has less energy than the absorbed light, resulting in a loss of energy. This value is constant over the entire EEM. Chapter 6 describes the pre-treatment methods used to correct for scattering effects.

Multivariate methods have been developed to quantify and classify components which act as fluorophores. Some common techniques include Parallel Factor Analysis (PARAFAC) and n-way partial least squares (NPLS) regression. These two methods are examined in detail in Chapter 6. Table 2-2 is a review of multivariate methods used to quantify and classify fluorescence EEMs for a variety of samples with a focus on those from a biological source.

Table 2-2: Review of multivariate methods used to quantify or classify fluorescence EEM spectra with a focus on largely biologically sourced samples.

Sample Type	Measured Component	Pre-treatment	Multivariate Method	Results	Reference
Carrot baby food	Neoformed compounds (e.g. furosin, furan)	First order Rayleigh replaced by missing values and estimated using maximum likelihood method	PARAFAC	• Furan is a carcinogen which has been found at higher levels than expected in processed vegetables • PARAFAC model was highly correlated to the measured compounds ($R^2 > 0.94$) • The neoformed compounds were found to be lower in purees from semi-frozen carrots than fresh and pasteurized materials	(Acharid *et al.* 2012)
Sherry vinegar	Classification on age	Rayleigh scattering removed with missing values	PARAFAC followed by PLS-DA, SVM	• Goal: successfully classify Sherry vinegar by age (6 months, 2 years, and 10 years) • PARAFAC model gave information about the fluorescent molecules and their relative amount • SVM was the most adequate classification method as there was almost no error in classifying the aged wine	(Callejón *et al.* 2012)
Graphene oxide (humic acid formation on surface)	Humic acid fractionation	Response of blank deducted from samples	PARAFAC	• PARAFAC identified two components that were humic-like in shape and location • PARAFAC modelling can be used to track changes in the molecular size of humic acid • One humic-like components was determined to be larger sized and had a greater adsorption affinity	(Lee *et al.* 2015)

Water	Dissolved organic matter (Suwanee River fulvic acid, Nordic Reservoir natural organic matter	Raman scatter correction Missing values for Rayleigh scattering Intensity value of 0 assigned to emission below excitation wavelengths	PARAFAC	• PARAFAC model revealed that there were 6 independent components • The components behave differently as the pH is adjusted which can be monitored with the model • PARAFAC was also used to characterize the components in the sample (e.g. humic-like, protein-like) • Overall, PARAFAC was determined to be a promising technique to characterize the functions of dissolved organic matter	(Yan *et al.* 2013)
Water	Dissolved organic matter	Response of blank deducted from samples Normalized using quinine sulfate units	PARAFAC	• Dissolved organic matter in drinking water can re-promote the growth of bacteria and biofilm causing corrosion • Maximum fluorescence intensities were used to represent component concentrations • The PARAFAC model was successful in relating kinetic rate to individual dissolved organic matter components • The model provided new insights on the underlying mechanisms related to the photodegradation	(Dinh & Hur 2015)
Honey	Classification	Missing values inserted for Rayleigh scattering	PARAFAC followed by PLS-DA	• The fluorescence data on honey was taken using a front-faced fluorescence approach • Goal: classify honey based on botanical origin and indent fake honey samples • 6 components were found in the PARAFAC model • PLS-DA can successfully detect fake honey samples with 100 % sensitivity and specificity	*(Dramic et al.* 2015)

Cell culture media	Tyrosine, tryptophan, folic acid, pyridoxine	Missing values inserted for Rayleigh scattering	PARAFAC, multivariate curve resolution, NPLS	• PARAFAC monitored the decrease of intrinsic media components and the photo-degradation products • NPLS could estimate changes in concentration for any product that was not extensively photo-degraded • Multivariate curve resolution was the more accurate method for quantifying degraded media	(Calvet *et al.* 2014)
Coffee	Classification	Rayleigh scattering was replaced by missing values	PARAFAC, NPLS-DA, unfolded PLS-DA	• Goal: geographically classify coffee • Unfolded PLS-DA performed better than NPLS-DA for the geographical classification of coffees • The f-scores for regions using unfolded PLS-DA method were greater than 0.8 for the training and test sets	(Botelho *et al.* 2017)
Cell culture media	Tryptophan, tyrosine	Rayleigh scattering was replaced by missing values	PARAFAC, NPLS	• PARAFAC was used to identify the fluorophores present in the samples • Both tryptophan and tyrosine were successfully quantified using NPLS with an accuracy of 4.5 and 5.5 %, respectively • It is possible to reduce the amount of error in prediction by reducing the range of the concentrations in the calibration set	(Calvet *et al.* 2012)
Cell culture media	Downstream product yield	Rayleigh scattering was replaced by missing values	NPLS-DA, NPLS	• NPLS-DA was used to accurately determine subtle composition changes that occur due to prolonged storage • The NPLS model was able to correlate small variances within the EEM to end product yield with an accuracy of ± 0.13 g/L	(Ryan *et al.* 2010)

2.5 Data Fusion

Data fusion is a process of combining several sources of data in order to find a uniform picture. This technique borrows ideas from diverse fields in order to combine the data into one model. The advantages of data fusion include improved detection, an increase in accuracy and reliability, and extending the available data (Khaleghi et al. 2013). Fluorescence and near-infrared spectroscopy both give information about the chemical composition of the sample and therefore an enhancement to the model can be anticipated by combining the spectral data.

Data fusion can occur at three different levels: low, medium, and high (Solano et al. 2012). Low data fusion combines data after separately pre-treating, medium data fusion extracts certain data features before combining, and high data fusion creates separate multivariate models where the outputs are combines.

Low level data fusion is the easiest and most commonly method used. However, it is important to note that each technique's dataset should be normalized and have about the same number of variables so that the final model does not depend more heavily on one technique than the others (Khaleghi *et al.* 2013).

Table 2-3 is a review of data fusion using various spectroscopic techniques to classify or quantify components in the spectra. Data fusion is not commonly used to combine NIR and fluorescence spectra. Data fusion is much more common between NMR and other spectroscopic techniques (Bro *et al.* 2013; Dearing *et al.* 2011; Fernández *et al.* 2013; Anibal *et al.* 2011).

Table 2-3: Review of data fused spectroscopic techniques used with multivariate methods to quantify or classify components.

Sample Type	Measured Component	Fused Data Sources	Multivariate Method	Results	Reference
Spices	Classification	UV-Visible spectra, NMR	PLS-DA	• Detecting banned Sudan dyes in commercial spices • Four fuzzy aggregation connective operators (minimum, maximum, product, and average) used to fuse the data and perform PLS-DA • Data fused model is more effective than the individual techniques (model correctly classified 80-100 % of the samples individually and 97-100 % combined)	(Anibal *et al.* 2011)
Soy hydrolysates	Integrate viable cell density (IVCD) and immunoglobulin G (IgG)	NIR, Raman, 2D fluorescence, X-ray fluorescence	Ensemble PLS (EPLS), multi-block PLS (MBPLS)	• Generated unified estimation from sub models of the multiple datasets (high level fusion) • Raman spectra yielded the best model individually • EPLS models outperformed MBPLS for the fused data • Data fused models using EPLS exhibit the best prediction accuracy overall • Certain predictions actually are better using the individual models, but the overall model is by far the best	(Lee *et al.* 2012)
Crude oil	Characterization of important parameters	Raman, IR, NMR	PLS	• Low level data fusion performed after individual pre-treatments • Data reduction of the NMR spectra before fusing • Data fusion successfully increased the model’s performance as seen by the significant reduction in the root mean squared error of prediction	(Dearing *et al*. 2011)

Azo-dyes (acid orange 61, acid red 97, acid brown 425)	Photodegradation process of the dyes	US-Vis spectroscopy with a diode array detector (DAD), NMR	Multivariate curve resolution-alternating least squares	• Low level data fusion with individual pre-treatment • After scaling both sets, the NMR dataset had to be reduced to have around the same number of data points as the UV-Vis • Fusion of data before multivariate analysis provides better results than the individual datasets (mostly with acid orange 61) • There was very little difference in the acid red 97 and acid brown 425 models between the UV-Vis and fused models	(Fernández *et al.* 2013)
Transformer insulating oil	Interfacial tension, colour	NIR, molecular fluorescence, NMR	PLS, variable importance in the projection (VIP) scores	• Separate pre-treatments were performed on each dataset before fusing • Compared full fused dataset as well as a VIP score reduced set (VIP scores evaluated as a means to compress the fused data) • The best prediction method for interfacial tension and colour used the fused dataset which had been compressed using the VIP scores	(Godinho *et al.* 2014)
Human plasma	Detection of colorectal cancer	Fluorescence, NMR	PCA, Area under the curve	• Area under the curve as the spectral variables to reduce the likeliness of overfitting (due to a limited amount of samples) • Low level data fusion after data reduction • The classification power improved with the fused dataset	(Bro *et al.* 2013)

Chapter 3

Materials and Methods

3.1 Flour Sources

Pin milled navy bean flour (Canadian International Grains institute, Winnipeg, Manitoba, Canada) and regular (hammer) milled navy bean flour (International Food Products, Chatham, Ontario, Canada) were used in the electrostatic separation process (Section 3.3) without modification aside from drying.

3.2 Chemical Sources

Armesco (Solon, Ohio, USA): acetic acid (glacial)

British Drug Houses (United Kingdoms): calcium chloride dehydrate (A.C.S. reagent); potassium sodium tartrate (A.C.S. reagent)

Ricca Chemical Company (Arlington, Texas, USA): Nessler reagent (R5250000)

Sigma Aldrich Chemical Company (St. Louis, Missouri, USA): ammonium sulfate; amyloglucosidase (≥ 250 U/mL, from *Aspergillus*); 3,5-dinitrosalicylic acid (DNS); ethanol, potassium sulfate (A.C.S reagent); selenium oxychloride; sodium hydroxide (powder, A.C.S reagent); sulfuric acid; thermostable α-amylase (> 20000 U/mL).

3.3 Electrostatic Separation

The laboratory scale triboelectrostatic separator (schematic shown in Figure 3-1) was designed by Advanced CERT Canada Inc. (Waterloo, Ontario, Canada) which was used for the dry fractionation of the navy bean flour. The flour was first dried at 70 ^{o}C for 24 hours before approximately 25 grams were placed into the fluidized bed. The air (pressurized at ~200 kPa) was passed through a drying column and introduced into the fluidized bed at a constant flowrate. The particles are suspended by a combination of a magnetic stirrer in the fluidized bed and high pressure dried air. The suspended particles then enter the tribocharger tube pneumatically.

Within the tribocharger the flour particles gain a charge resulting from particle-wall and particle-particle collisions. The tribocharger was made of a polytetrafluoroethylene (PTFE) tube with 3/16 in. (4.76 mm) outside diameter and varied in length and shape (straight or coiled). PTFE was chosen because it was most effective in positively charging the protein- and carbohydrate-rich particles of the navy bean flour based on the its work function (see Section 2.1.1 for background).

The actual separation of the protein- and carbohydrate-rich particles takes place in the rectangular separator unit. Within the unit a 66 by 25.5 cm copper plate is negatively charged by a high voltage DC power supply source. When the positively charged particles enter the separator they interact with the electrostatic field created by the copper plate and fractionate along the bottom of the separator and on the copper plate. The protein-enriched fractions end up on the copper plate while the carbohydrate-enriched fractions end up on the bottom of the separator.

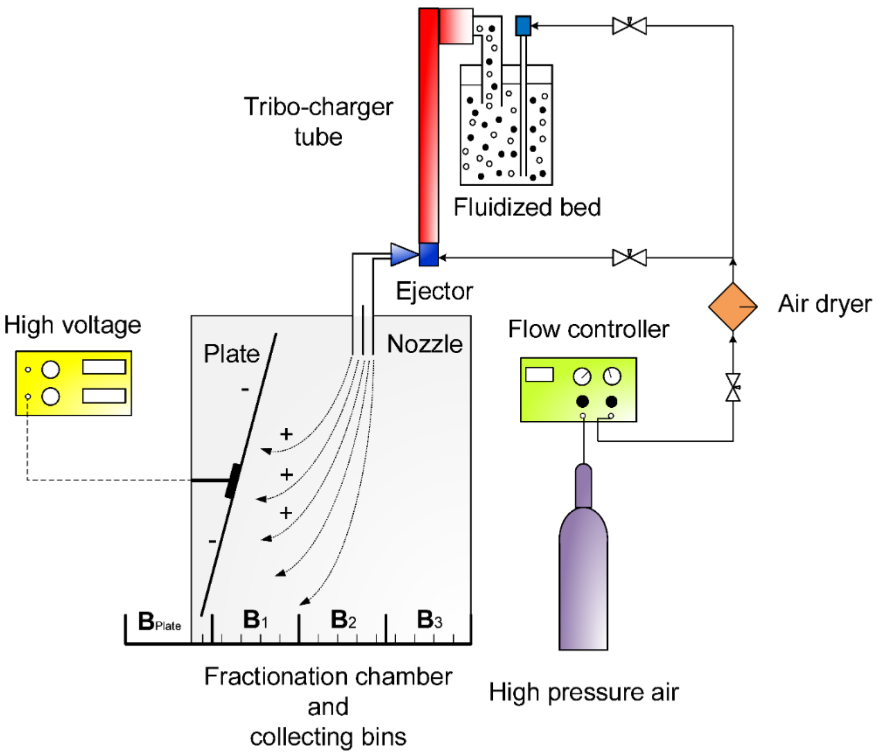


Figure 3-1: Schematic of the electrostatic separation **(**from Tabtabaei *et al*. 2016).

3.3.1 Fraction Collection during Electrostatic Separation

Seven fractions of flour were collected after the completion of the batch electrostatic separation. The fraction locations are shown in Figure 3-2 while a comparison of the areas is shown in Table 1. The bottom of the separator and the plate were divided in to three sections, (B_1, B_2, and B_3) and (P_B, P_M, P_T), respectively. . A sixth fraction was material collected from the side of the separator (S). When collecting the faction they were collected in the following order: B_1, B_2, B_3, S, P_B, P_M, and lastly P_T.

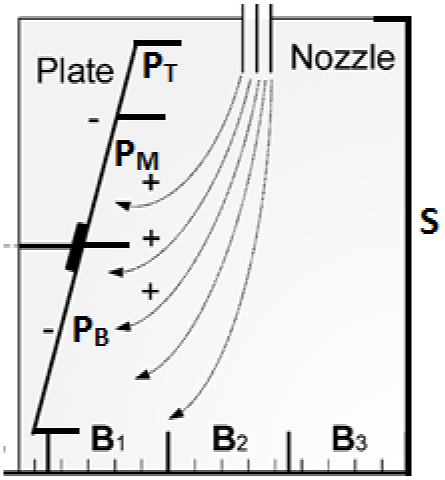


Figure 3-2: Electrostatic separator fractions (adapted from Tabtabaei *et al*. 2016).

Table 3-1: Fraction areas for the electrostatic separator.

Fraction	**Area (cm^2)**
B_1	585
B_2	468
B_3	408
P_B	841.5
P_M	382.5
P_T	459
S	3927

3.4 Protein Content Determination via Kjeldahl Digestion

The protein content of all the navy bean flour fractions was determined by the mircrodetermination of Kjeldahl nitrogen (Lang 1958).

3.4.1 Digestion Solution Preparation

40 grams of potassium sulfate was added to 250 mL of milli-Q water. The mixture was then placed on a magnetic stirrer and positioned in an ice water bath. 250 mL of sulfuric acid was added to the mixture. The sulfuric acid was added slowly while monitoring the temperature of the ice bath. Lastly, 2 mL of selenium oxychloride was added. The digestion solution was mixed for an additional 2 h.

3.4.2 Kjeldahl Standard Solution Preparation

Ammonium sulfate was added to distilled water at a concentration of 4.714 g/L. 2 mL of the standard solution was placed in a 30 mL Kjeldahl flask and the same procedure followed as for the samples.

3.4.3 Acid Digestion

46-50 mg of flour (record weight) was added to a 30 mL Kjeldahl flask and 5 mL of digestion solution was added. The Kjeldahl flask was then gently heated at low for 30 min. The temperature was then increased to a setting of 2 for 1.5 h. After digestion, the samples were cooled for 30 min at room temperature. 25 mL of distilled water was added to each sample and the new volume recorded.

3.4.4 Nesslerization

Each sample was added in quintuplicate to a clear flat bottom well plate with water and Nessler reagent in the amounts given in Table 3-2. Flour fractions with a higher protein content ($P_{B,}$ P_M, and P_T) were diluted two fold. The Kjeldahl standard had multiple dilutions (1, 2 4, 8 times) in the plate to create the calibration curve. A blank was also prepared consisting of only Nessler reagent and water.

Table 3-2: Kjeldahl dilution amounts.

Dilution	**Water (µL)**	**Sample (µL)**	**Nessler reagent (µL)**
1	200	40	50
2	220	20	50
4	230	10	50
8	235	5	50
Blank	240	0	50

3.4.5 Protein Content Measurement

After the plate contents had been well mixed, the resulting colour was measured at 420 nm using a spectrophotometer (Bio Tek Instruments, USA). The protein content of the sample was then calculated using equation (1). A sample calibration curve can be found in Appendix A. All protein content measurements were done in duplicate.

$$PC = \frac{(A - B) * k * V * 6.25}{SV * w * 1000} * 100\% \quad (1)$$

where: PC: protein content (%)

A: sample absorbance

B: blank absorbance

k: standard calibration curve slope

V: volume (measured, mL)

6.25: factor converting crude nitrogen to protein (Iban 2003; Ndiowere 1984; Braaksma 1995)

SV: sample volume in plate well (mL)

w: mass of sample (measured, mg)

3.5 Starch Determination using the Dinitrosalicylic (DNS) Acid Assay

The starch hydrolysis procedure was adapted from the extended AOAC procedure for starch in cereal grains published by Hall (2009) and as outlined in the Megazyme Total Starch Assay test-kit (Megazyme International Ltd.). Table 3-3 provides details for the preparation and storage of the solutions required for the assay.

Table 3-3: DNS solution preparation and storage

Reagent	**Preparation and Storage**
Acetate Buffer	Add 5.8 mL of glacial acetic acid (1.05 g/mL) to 900 mL of distilled water. Adjust the pH to 5.0 by adding approximately 30 mL of 1 M sodium hydroxide solution. Add 0.74 g of calcium chloride dihydrate and dissolve. Adjust the volume to 1 L with distilled water and store the buffer at 4 °C. The solution should be stable for 2 months at 4 °C.
Thermostable α-amylase solution	Dilute the volume of amylase solution containing 300 units of protein (i.e. 1.5 μL) by a factor 1:30 with acetate buffer. This solution will suffice for 10 assays and can be scaled up for larger volumes. When not in use, the solution should be frozen in propylene tubes in 3 mL- sized aliquots.
Amyloglucosidase from *Aspergillus niger*	Obtain amyloglucosidase solution (>200 units/mL) and directly dispense 20 units (i.e. 100 μL) into each sample tube. Store the enzyme at 2 to 8 °C.

Some modifications were made to the method. First, a 15 mL polypropylene centrifuge tube was used in place of glass test tubes to avoid the use of glass during the centrifugation step. The incubation times with α-amylase were increased from 6 to 12 minutes (Megazyme assay test-kit manual 2009). Hall (2009) calls for a second ethanol wash. Since the navy bean flour used in this work contains low levels of free sugars and lipids, the second ethanol wash was not found to be necessary for reliable results. Lastly, a standard followed the complete assay protocol (from the starch hydrolysis). Therefore, enzymes or ethanol traces from the procedure could not contribute to the glucose measurement during the DNS assay (Numan 2015).

The procedure contains four steps: removal of free sugars and lipids, starch hydrolysis, glucose determination, and data analysis.

3.5.1 Removal of Free Sugars and Lipids

50 mg of solid sample was added to a 15 mL polypropylene centrifuge tube. A separate empty centrifuge tube was used as the standard and contained no sample. 5 mL of 80 % v/v ethanol was added to the sample and incubated for 5 min at 80-85 ^{o}C. The sample was then vortexed and an additional 5 mL of 80 % v /v ethanol added. Next, the sample was centrifuged at 3750 x*g* rpm for 10 min.

3.5.2 Starch Hydrolysis

3 mL of buffered α-amylase enzyme solution (Table 3-3) was added to the sample which was then incubated in a boiling water bath for 12 min. At 4, 8, and 12 minute marks the sample was vortexed vigorously to ensure the flour was well-distributed. The centrifuge tubes were placed in a 50 ^{o}C water bath for 5 min. 100 μL of amyloglucosidase was added; the sample was then vortexed and incubated at 50 ^{o}C for 30 min.

6.5 mL of distilled water was added and the total volume of the sample recorded. The sample was then centrifuged at 3750 x*g* for 10 min. 1 mL of supernatant was collected and added to 9 mL of milli-Q water. The mixture is once again vortexed and a 1 mL aliquot added to a new 15 mL polypropylene centrifuge tube.

3.5.3 Glucose Determination

2 mL of DNS solution was added to the aliquot. A glucose standard was prepared (1 mg/mL in milli-Q water) and diluted to concentrations of 0.1, 0.2, 0.3, and 0.4 mg/mL. 1 mL of milli-Q water was used as the blank for the glucose standard. 2 mL of DNS solution was also added to the glucose standards and blank. All samples and standards were then incubated for 5

min in a boiling water bath. Samples were taken out and cooled for 30 minutes at room temperature. 9 mL of milli-Q water was added to all the samples, standards, and blanks. 3 mL aliquots of each sample, standard and blank was transferred to a glass cuvette. The absorbance was read using a spectrophotometer (Spectronic, Genesys 5) at 540 nm and recorded. The standards were read against the distilled water blank while the samples were read against the standard with no sample.

3.5.4 Data Analysis

The starch content of the samples was calculated using equation (2). A sample standard curve can be found in Appendix A. All starch content measurements were done in duplicate.

$$SC = (m * A + b) * DV * \frac{100}{w} * \frac{162}{180} \quad (2)$$

where: SC: starch content (%)

m: slope of the glucose calibration curve

A: absorbance of sample (measured)

b: y-intercept of the glucose standard curve

DV: final sample volume (mL) = 10 * recorded volume (section 3.5.2)

w: mass of sample (measured, mg)

$\frac{162}{180}$: adjustment from D-glucose to anhydro D-glucose (as found in starch)

3.6 Near Infrared (NIR) Spectroscopy Set-up

Approximately 72 mg of flour was compressed (less than 3 mm in thickness) into a sample holder (3.5 cm diameter) with a plunger as shown in Figure 3-3. The compressed flour was inserted into the near infrared spectrophotometer (Perkin Elmer, Lambda 750S) and analyzed from 1200 to 2500 nm using an integrating sphere. All NIR scans were the result of averaging two scans. There were 102 flour samples that were analyzed for absorbance. The samples were split into sets of 10 where a standard sample was run before each set.

The following settings were used for the NIR spectrophotometer: The slit width was set to 2 nm. The PMT response, which defines the average signal time, was set to 0.2 s. Lastly, the number of cycles was set to 2 as two scans were averaged for the final sample spectra.

The spectrophotometer was zeroed before each set of 10 samples. The zeroing was done through the spectrophotometer's autozero program with no sample inserted. Following the autozero, a standard whey protein isolate (~97 % protein) was run under the same conditions. This standard was run before each set of 10 samples to account for the day to day variance for the spectrophotometer. The results for the standard can be found in Appendix B.

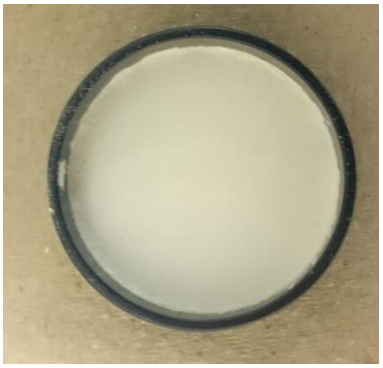


Figure 3-3: Compressed flour sample.

3.7 Fluorescence Spectroscopy Set-up

The same compressed flour sample (Section 3.6) was analyzed for its fluorescence excitation-emission matrix (FEEM). The FEEM was acquired using a spectrofluorometer (Agilent Technologies, Cary Eclipse Fluorescence Spectrophotometer) equipped with a front-faced fluorescence fibre-optic probe. The set-up used for acquiring the FEEM is shown in Figure 3-4. The samples were adjusted such that the probe's distance to the sample was as close as possible without making direct contact with the sample. This height was at the largest scaling marker (see 'height setting' on Figure 3-4).

The samples were analyzed with excitation wavelengths varying from 250 to 380 nm (increasing by 10 nm) and emission wavelengths between 300 and 600 nm (increasing by 1 nm) using a fibre optic probe at an angle of 45°. All fluorescence spectra analyzed were the result of averaging two scans. The PMT voltage was set at 680 V. The slit widths for excitation and emission were both set at 5 nm. There were 102 samples tested for fluorescence and were also split into sets of 10 where a standard was run before each set.

The fluorometer was zeroed before each sample using a blank sample holder set to the same distance from the probe. Before each set of 10 samples, a standard of whey protein isolate (~97 % protein) was analyzed under the same conditions as the samples to test for day to day variance from the spectrofluorometer. The results for the standard can be found in Appendix B.

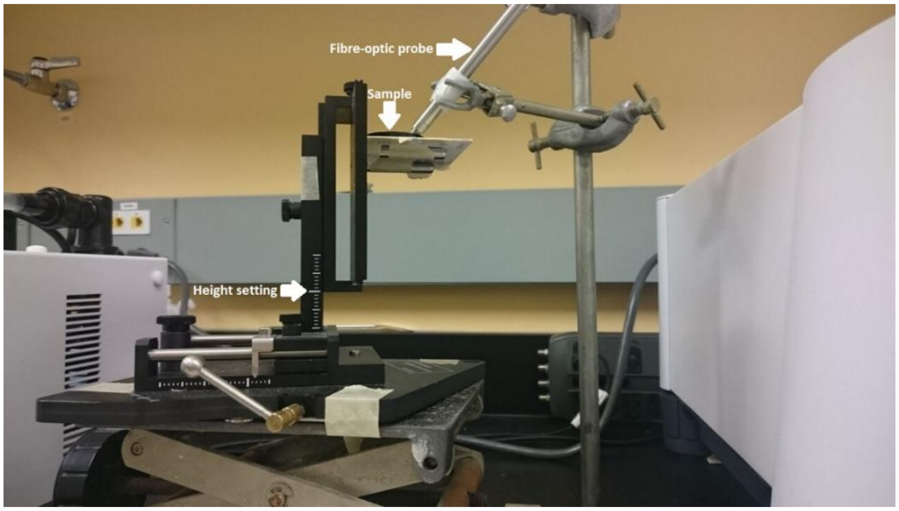


Figure 3-4: Front faced fluorescence set-up.

3.8 Aerodynamic Particle Size Analysis

The volume- and number-weighted distribution curves were obtained using a TSI 3603 (TSI Incorporated, Shoreview, MN, USA) aerodynamic particle sizer. This was done by Grace Li at the University of Western Ontario.

Aerodynamic particle sizers use the concept of inertia to size particles. Particles and air flow is constructed using a nozzle which accelerates the flow. The particles are also accelerated but at different rates which depends on the particle surface area and mass. This method assumes that the particles are spherical and have unity density.

The velocity of the particles is measured from the particles passing through two laser beams. As the particles pass through the light, they produce a pulse of scattered light. The particle sizer has an elliptical mirror which collects the scattered light onto a photodetector. Therefore, the time delay between the pulses is relates to the particle velocity which in turn is related to the particle diameter.

Chapter 4

Effect of Milling Techniques on Electrostatic Separation

4.1 Particle Size of Pin and Regular Milled Navy Bean Flour

As discussed in Chapter 2, particle size is an important factor for both triboelectric charging and electrostatic separation. Different milling techniques yield differing particle size distributions and therefore will have an effect on separation. Two different milling techniques for navy bean flour were evaluated (regular and pin milled) using the electrostatic separation method described in Section 3.3.

Pin milling works by introducing the feed product onto various spinning rotors with different configurations of pins or blocks that act as impactors (Rajkovich 2017). Regular milling, or ball milling, occurs due to the impacts, compression, and grinding between the sample and the grinding media (Shin *et al.* 2016).

Figure 4-1 presents the particle size distributions of pin and regular milled navy bean flour. Both navy bean flours have a protein content of just over 27% so they have very similar compositions even though they are from different sources. Pin milling the navy bean yields a much smaller average particle size, 5.98 compared to 16.55 μm, which is almost three times as small. Also, over 80% of the particles for pin milled flour have a diameter under 10 μm while the regular milled flour only about 21% under 10 μm.

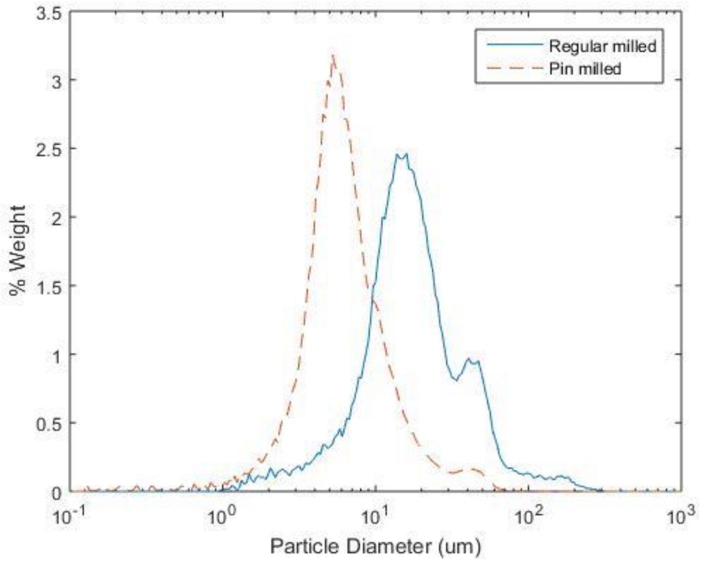


Figure 4-1: Particle size distribution of pin and regular milled navy bean flour.

4.2 Optimized Separation Conditions

Since the particle size distributions of pin and regular milled flour are different, it is possible that the optimum conditions for the electrostatic separation would also differ. Optimization was done similar to Tabtabaei *et al.* (2016) using mixed level full factorial experiment where the main focus was to optimize protein content and extraction of the protein-enriched fraction. The protein content was measured using Kjeldahl digestion (Section 3.4). The extraction of protein is the percentage of total protein in the protein-enriched fraction. Table 4-1 shows the optimized separation parameters for pin and regular milled navy bean flour as well as the results for protein content and extraction of the protein-enriched fraction

Table 4-1: Optimal parameters and results for pin and regular milled navy bean flour (pin milled results from Tabtabaei *et al.* (2016)).

	Voltage (kV)	Flowrate (LPM)	Length (cm)	Protein Content (%)	Protein Extraction (%)
Pin Milled	-5	7	60	40.7	44.8
Regular Milled	-1	6.1	250	32.5	78.3

Since the particles of regular milled bean flour are larger, it takes more collisions for the particles to achieve the same amount of charge per unit volume. Therefore, the length of the triboelectric tube was increased. This resulted in more collisions that needed to be offset by a slower flowrate and lower voltage on separator plate.

It was not possible to obtain a protein content of the protein-enriched fraction comparable to the same level regular milled flour. The regular milled flour had a much higher average particle diameter and pin milling inherently created particles that had more protein content. Instead, a much higher extraction for the protein-enriched fraction was achieved but at a lower protein content.

4.3 Particle Size Distribution of Navy Bean Flour Fractions

After separation, the flour fractions were divided into two larger fractions: protein- and starch-enriched. Figure 4-2 and 4-3 shows the particle size distributions for the optimal separations of pin and regular milled flour, respectively.

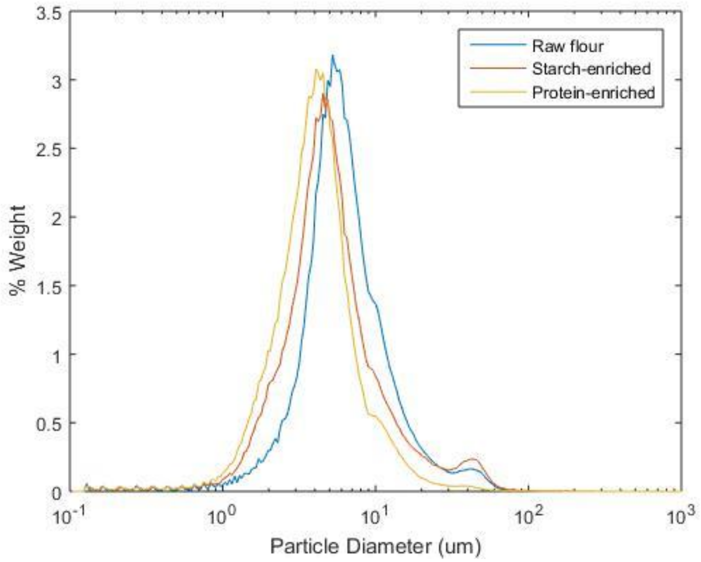


Figure 4-2: Particle distribution of the fractions for pin milled navy bean flour under optimal conditions.

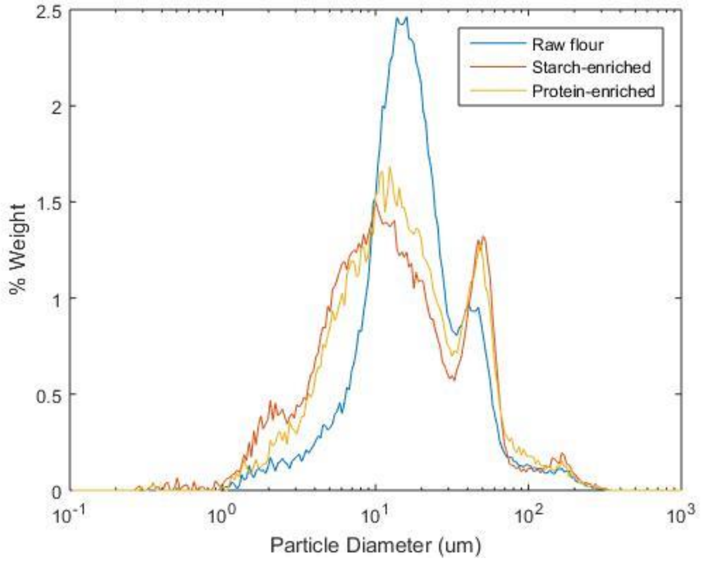


Figure 4-3: Particle distribution of the fractions for regular milled navy bean flour under optimal conditions.

There is not much difference between the particle size distributions for the enriched fractions and the raw pin milled navy bean flour. However, the regular milled flour particle distribution is much different than the fractions. This difference occurs mainly in the 10-40 μm particle size range. It may be that there exists aggregates of starch and protein particles that disaggregate when they become charged in the triboelectric charger since there is a higher percentage of particles for the separation fractions in the 1-10 μm range.

Table 4-2 shows the protein content and average particle diameter for the enriched fractions for pin and regular milled navy bean flour under optimal separation conditions. As expected, the average particle size for the regular milled flour is much higher than the pin milled flour. Also, the raw navy bean protein content for both milling techniques is very similar. Lastly, the protein content of the pin milled protein-enriched fraction is 7 % higher than the protein-enriched fraction for the regular milled flour.

Table 4-2: Protein content and average particle diameter for raw and enriched fractions under optimal conditions for pin and regular milled navy bean flour (pin milled results from Tabtabaei *et al.* (2016)).

	Protein Content (%)		Average particle diameter (um)	
	Pin Milled	Regular Milled	Pin Milled	Regular Milled
Raw Flour	26.84 ± 1.1	27.63 ± 0.73	5.98	16.55
Protein-enriched	42.90	35.23	4.12	14.70
Starch-enriched	18.53	16.30	4.78	12.93

Chapter 5

NIR Results

5.1 Navy Bean Flour Samples

The samples used in the NIR model are navy bean flour fractions from the electrostatic separation and the raw navy bean flour. There are 102 samples that are split into calibration (82 samples) and validation (20 samples) sets. In the calibration set, the samples are further split into fractions: protein-enriched, starch-enriched, and raw flour. The protein-enriched samples are collected from the electrostatic plate while the starch-enriched fractions are collected from the bottom of the electrostatic separator. The schematic of the collection can be found in Section 3.3.

5.2 Data Pre-treatment

5.2.1 Multiplicative Scatter Correction (MSC)

MSC is a very common pre-treatment method for NIR spectra (Rinnan *et al.* 2009). It gives an estimation for the scatter of each sample relative to the scatter of an ideal sample. The ideal sample or reference spectrum is usually taken as the average spectrum for of all the samples. This technique corrects for baseline shifts by correcting each sample's scattering to the same level.

The algorithm for MSC (Geladi *et al.* 1985) is shown below. The first step of the pre-treatment process is to mean center the sample and reference spectra.

$$x_c = x - \bar{x} \quad (1)$$

$$r_c = r - \bar{r} \quad (2)$$

The next step is to determine the scaling coefficient (b) and correct for the baseline shift. The final pre-treated sample spectrum is depicted by x_{msc}.

$$b = (r_c^T * r_c)^{-1} * r_c^T * x_c \quad (3)$$

$$x_{msc} = \frac{x_c}{b} + r_c \quad (4)$$

5.2.2 Standard Normal Variate (SNV)

SNV has the same basic format as MSC and usually leads to very similar results (Kamruzzaman & Makino 2015; Vázquez 2014). The difference is that each spectrum is processed on its own. The spectrum is centered by its mean and scaled by its standard deviation (as shown in Equation 5) which removes the baseline shift between data samples. This pre-treatment technique is sensitive to noisy entries in the spectra, but is less sensitive to outliers in the data since each spectrum is treated separately.

$$x_{snv} = \frac{x_{i,j} - \overline{x_\iota}}{\sigma_i} \quad (5)$$

The mean and standard deviation of each spectra is calculated by equations 6 and 7, respectively.

$$\overline{x_\iota} = \frac{\sum_{j=1}^{n} x_{i,j}}{n} \quad (6)$$

$$\sigma_i = \frac{\sqrt{\sum_{j=1}^{n}(x_{i,j} - \overline{x_\iota})^2}}{(n-1)} \quad (7)$$

5.2.3 Savitzky-Golay (SG) Differentiation

SG derivation (Savitzky *et al.* 1964) involves fitting a polynomial to the raw data. The polynomial increases the signal to noise ratio without distorting the data. Also, by differentiating the data, the baseline shift is removed. The polynomial can have a minimum order of the derivative number (e.g. linear for first derivative) and the derivative is taken over a moving window. The moving window is the number of points used to calculate the polynomial. A 19 point moving window was used to determine the first derivatives of the navy bean flour and fraction samples in this work.

5.3 Multivariate Methods

5.3.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a mathematical procedure for resolving sets of data into orthogonal components whose linear components approximate the original data. This technique adequately describes the data using far fewer factors with no significant loss of data (Burns 2008). The principal components are given in the order of increasing variance explained in the data set. That is the sample variances of the given points with respect to those derived coordinates are in decreasing order.

The covariance matrix is needed to determine the principal components of the n x p matrix X, as shown in Equation 8.

$$var(X) = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix} = W \qquad (8)$$

The coefficients of the principal components are the eigenvectors of the covariance matrix. Therefore, matrix X can have a total of p principal components. Equations 9 and 10 show the determination of eigenvalues (λ) and eigenvectors (v), respectively. The principal components are then arranged in order of decreasing variance explained in the matrix X. Therefore, the principal components that explain the highest variance in the dataset are shown first.

$$\det(W - \lambda I) = 0 \quad (9)$$

$$(W - \lambda I)v = 0 \quad (10)$$

The eigenvectors (v) are also called loadings. The principal component scores (S) are determined by the cross product of the principal component loadings and the dataset as shown in Equation 4. The scores are a linear combination of the original dataset. The calculation of principal component scores decreases the amount of variables to the number of principal components used.

$$S = v * X \quad (11)$$

In summary, principal components are the result of applying a mathematical technique that generates eigenvalues. The largest eigenvalues are used from a covariance matrix of the dataset where the components are orthogonal and linear combinations of the data. The eigenvalues are put in an order of increasing variance explained in the original dataset. Therefore, choosing a number of principal components reduces the dataset to the same number of variables which give the highest amount of variance explained in the dataset.

5.3.2 Cross Validation

The k-fold cross validation method is used to evaluate the number of principal components to be used in the model. In this method, the dataset is split into k subsets. Each subset is then used as the validation set once while the remaining subsets create the calibration model. The advantage to the k-fold cross validation is that it allows the entire calibration set to be used to test and train the model. 10-fold cross validation is used in this study.

The final model is then evaluated using the root mean squared error of calibration (RMSEC) and root mean squared error of cross validation (RMSECV) shown in Equations 12 and

13. The equations have the same form with the only difference being the RMSECV only uses the error found when the sample is included in the validation set.

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{N}(y_{p,i} - y_{m,i})^2}{N}} \quad (12)$$

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{N}(y_{pv,i} - y_{m,i})^2}{N}} \quad (13)$$

where: y_p: y-value predicted from the model

y_m: y-value from offline measurements

N: total number of samples

y_{pv}: y-value predicted using the model (only from the validation set)

The lower the RMSEC and RMSECV values are, the better the model. To avoid overfitting, the number of principal components is identified to be when the RMSECV curve plateaus as adding additional components will only train the model to the calibration set which makes the model less robust.

5.3.3 Principal Components Regression (PCR)

Principal components regression finds the relationship between the x-scores (S) and the offline measurements found in Y (i.e. protein content, starch content). The method involves solving for regression coefficient (b) which correlate the independent and dependent variables, Y and X, respectively while minimizing the least square errors.

$$Y = bX \quad (14)$$

PCR requires the x-variable scores from PCA then projects Y onto the scores solving for the least squares (A) as given by Equation 15.

$$A = (S' * S)^{-1} * S' * Y \quad (15)$$

The matrix A can then be converted into the regression coefficients by combining with the x-variable loadings also found in PCA as seen in Equation 16.

$$b = v * A \quad (16)$$

The disadvantage of using this method is that the principal components do not take into account the independent variables (Y) when determining the scores and loadings of X.

5.3.4 Partial Least Squares Regression (PLS)

Partial least squares regression finds the best linear relationship between the principal components of the NIR spectra and offline measurements (in this study protein and starch content). This method attempts to find a regression coefficient (b) correlating the dependent and independent variables while minimizing the variance between the variables. PLS regression has the same final equation as PCR (Equation 17), however the regression coefficients will differ.

$$Y = bX \quad (17)$$

PLS is a bilinear regression model. Principal components are obtained differently from the PCA method described above. PLS regression attempts to find factors that maximize the amount of variance in X that is relevant for predicting Y.

The default method for PLS regression in the PLS Toolbox is the SIMPLS algorithm. It is derived to solve a specific objective function which maximizes covariance. It is faster than the NIPALS algorithm and involves calculating the weights, loadings, and scores of the matrices. The algorithm is shown below (Jong 1993) where X is an n x p and Y is an n x m matrix. The first step of the algorithm is pre-treatment which involves mean centering Y. Next, equations 20-35, the scores (T for X, U for Y) and loadings (P for X, Q for Y) of X and Y are solved for iteratively. The last steps involve solving for the regression coefficient, leverages, and variances shown in equations 36-39.

$$Y_o = Y - Mean(Y) \quad (18)$$

$$S = X' * Y_o \quad (19)$$

For $a = 1\ to\ A$ (usually between 1 and 1000)

$$q = dominant\ vector\ of\ S' * S \quad (20)$$

$$r = S * q \quad (21)$$

$$t = X * r \quad (22)$$

$$t = t - Mean(t) \quad (23)$$

$$normt = \sqrt{t' * t} \quad (24)$$

$$t = \frac{t}{normt} \quad (25)$$

$$r = \frac{r}{normt} \quad (26)$$

$$p = X' * t \quad (27)$$

$$q = Y_o' * t \quad (28)$$

$$u = Y_o * q \quad (29)$$

$$v = p \quad (30)$$

If $a > 1$

$$v = v - V * (V' * p) \quad (31)$$

$$u = u - T * (T' * u) \quad (32)$$

End if

$$v = \frac{v}{\sqrt{v' * v}} \quad (33)$$

$$S = S - v * (v' * S) \quad (34)$$

$$t = T, p = P, q = Q \quad (35)$$

Next a

$$B = R * Q' \quad (36)$$

$$h = DIAG(T * T') * \frac{1}{n} \quad (37)$$

$$varX = \frac{DIAG(P' * P)}{n - 1} \quad (38)$$

$$varY = \frac{DIAG(Q' * Q)}{n - 1} \quad (39)$$

5.4 Qualitative NIR Spectra Analysis

5.4.1 Analysis of Protein and Starch Peak Locations

The near infrared spectrum consists of wavelengths between 780 and 2500 nm (Workman & Weyer 2012). Molecular spectra result from vibrational energies caused by the absorption of infrared light. In this region, functional groups that include hydrogen absorb the wavelengths to vibrate which creates the spectra. Chemical bonds vary in strength and therefore the amount of energy it takes for the bond to vibrate. The different energies correspond to particular wavelengths where peaks in the spectra can be observed.

The goal in this work is to use the NIR spectra of powder specimens to quantify the concentrations of starch and protein in navy bean flour fractions that have been generated using a turboelectric-based separation approach. There are a number of peaks found in this range that pertain to protein and starch which are summarized in Table 5-1. It is possible for peaks to overlap when a sample has peaks in the same area. Therefore, multivariate methods must be used to view the spectra quantitatively. Figure 5-1 shows the spectra of potato starch and whey protein isolate in the near infrared range with the peaks found in Table 5-1 labelled on the curves. These are the regions where the greatest variance between different navy bean fractions should be observed.

Table 5-1: Summary of protein and starch related peaks in the NIR spectral range (compiled from Workman & Weyer 2012).

Number	Peak Location (nm)	Component	Bond Vibrating
1	1450	Starch	O-H polymeric
2	1500	Protein	Amide from protein
3	1540	Starch	O-H polymeric
4	1735	Protein	C-H Methyl C-H, amine associated
5	1935	Protein	O-H, amine associated
6	1960	Starch	O-H polymeric from polysaccharides
7	2050	Protein	N-H from amino acids
8	2100	Starch	C=O-O polymeric from glucose polysaccharides
9	2180	Protein	N-H secondary amides from protein

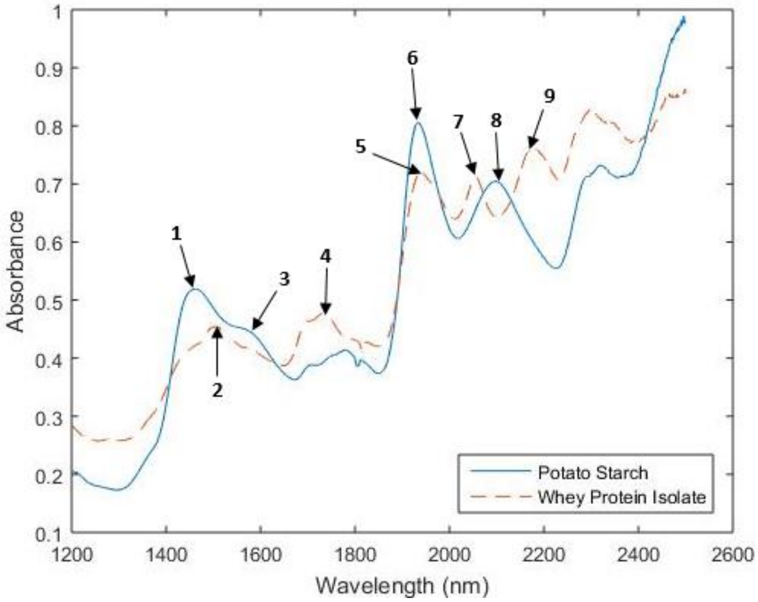


Figure 5-1: Near infrared spectra of potato starch and whey protein isolate with peak numbers corresponding to functional vibrations found in Table 5-1.

5.4.2 Raw Navy Bean Flour Fraction Spectra

Figure 5-2 shows the spectra of the original navy bean flour and protein- and starch-enriched fractions from triboelectric purification. There is a notable baseline shift between the samples. This is because of the path length of the light reflecting from the sample to the integrated sphere. As the path length between the sample and sphere increases, the baseline shift tends to be greater. The path length would increase as the thickness of the compressed pellet decreases. This is very hard to control as the compressed pellet flour samples are <3 mm in thickness.

There is a correlation between the fractions and the observed baseline shift which is a function of the particle size. This is the cause of a greater baseline shift shown in the raw spectra. The same mass is compressed into a pellet for analysis and therefore the greater the average particle size, the greater the distance between the sample and integrated sphere. Therefore, the starch-enriched fractions are compressed into a pellet with a smaller thickness than the protein-enriched fractions.

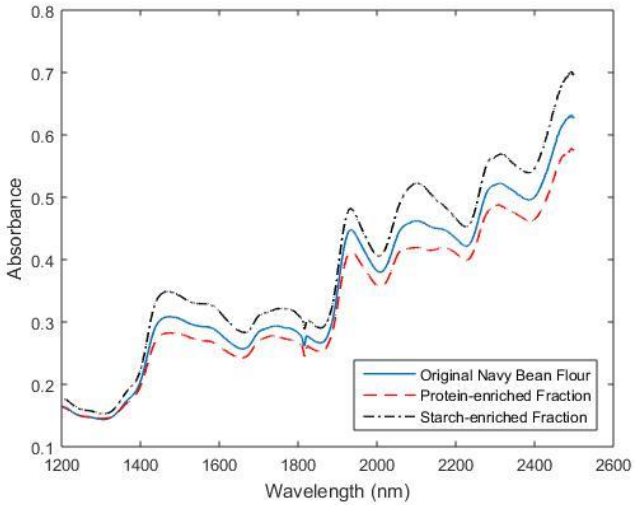


Figure 5-2: RAW spectra of original flour, a protein-enriched sample, and a starch-enriched sample.

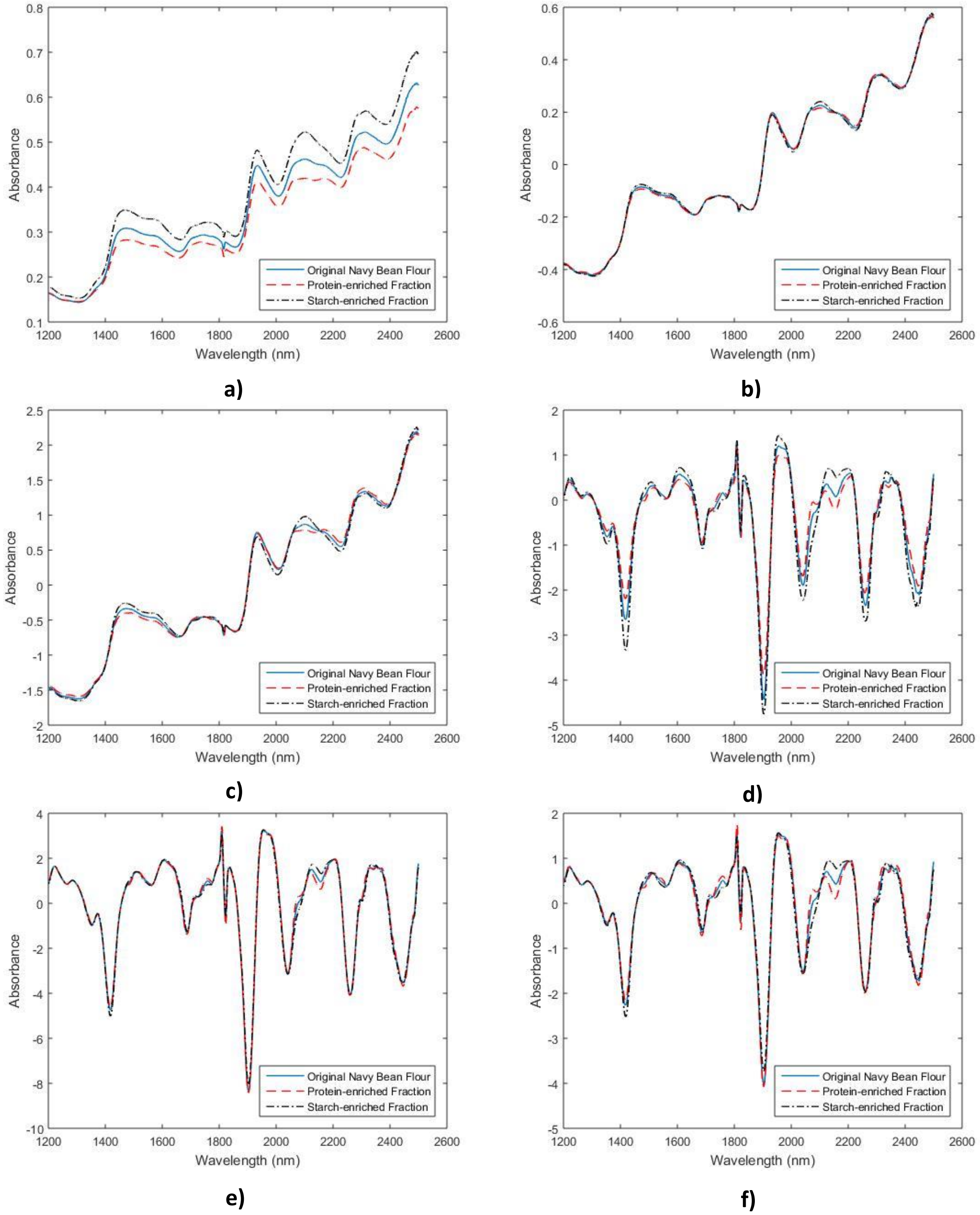
The variance between the actual curves of the spectra is more easily seen when the data has been pre-treated and will be discussed in Section 5.4.3.

5.4.3 Pre-treatment Analysis

As shown in Figure 5-2, there is a noticeable baseline shift between the samples. The baseline shift is also known to have a negative effect on the multivariate model. Therefore, various pre-treatments can be implemented to negate the effects of the baseline shift. Figure 5-3 compares the raw data of three samples (a protein-enriched fraction, a starch-enriched fraction, and original navy bean flour) to various methods of spectral pre-treatment.

After pre-treating the data, it is much easier to see where the differences lie between samples. The biggest variance occurs between 2000 and 2200 nm. In this area there are three overlapping peaks corresponding to protein and starch. The shape of the combination depends on the composition of the sample. This is best seen in the SNV pre-treated data (Figure 5-3c). The protein-enriched fractions have two very distinct peaks between 2000 and 2200 nm while the starch-enriched fractions have one distinct peak.

When comparing SNV and MSC pre-treatments (Figure 5-3b,c), it is worth noting the variance between the sample's spectra is much greater for the SNV pre-treatment. The three pre-treatments which include first derivative (Figure 5-3d-f) also show the greatest variance in the 2000-2200 nm range. The first derivative was also combined with MSC and SNV pre-treatment. In these cases, the first derivative of the spectrum was taken before the other pre-treatments were applied which shows that the first derivative did not entirely remove the baseline shift.



a)

b)

c)

d)

e)

f)

Figure 5-3: Various pre-treatment methods on the spectra of original flour, protein-enriched, and starch-enriched samples [a: RAW; b: MSC; c: SNV; d: first derivative (1D); e: 1D and MSC; f: 1D and SNV].

5.5 NIR Multivariate Model Results

The best model (from Table 5-4 and 5-5 in Section 5.6) was found using MSC pre-treatment and PLS. Therefore the analysis of steps found in the following sections will be done using this model. Section 5.6 compares the different pre-treatment effects as well as the differences between the PLS and PCR models.

5.5.1 Principal Components

The first step of using PLS is to determine the ideal number of components without overtraining the model to the calibration set. Overtraining leads to a less robust model and therefore will not perform as well with samples outside the calibration set. The number of components can be determined by using cross validation. Figure 5-4 shows the cross validation results for the RMSECV and RMSEC for starch and protein content using principal components between 1 and 10.

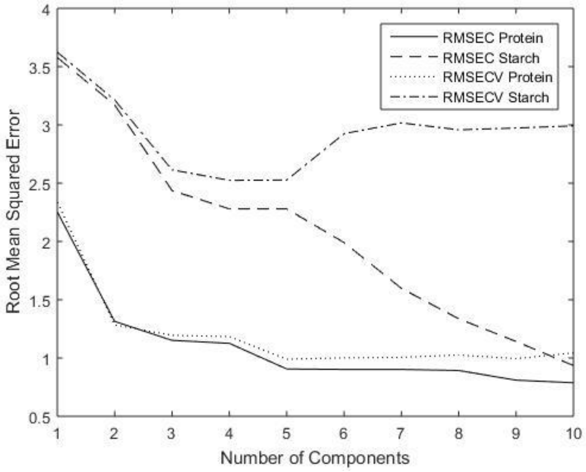


Figure 5-4: RMSECV and RMSEC for protein and starch content using 1 to 10 principal components with MSC pre-treatment and PLS to create the model.

The RMSECV plateaus when using more than 3 principal components which makes the ideal number of principal components 3. As shown by the RMSEC and RMSECV for starch content, increasing the number of components makes the model less robust. Even though the calibration error (RMSEC) continues to decrease, the cross validation error (RMSECV) suddenly increases when using more than 3 components.

Another way of determining the number of components is to examine the amount of percentage of total variance a certain number of components explains. This is shown in Figure 5-5.

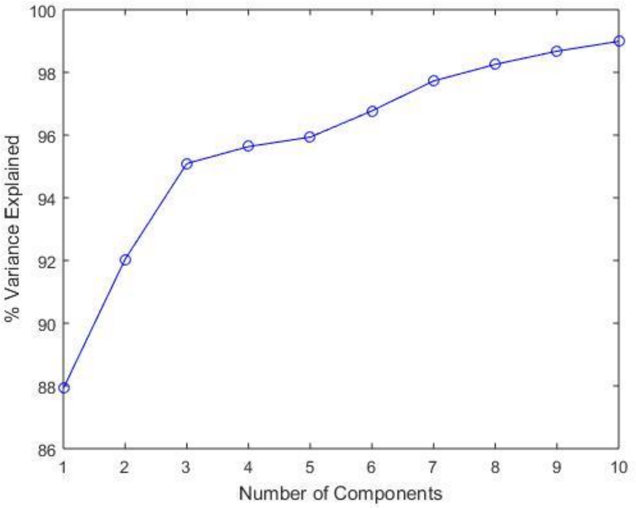


Figure 5-5: Percentage of total variance explained depending on the number of components used in the PLS model with MSC pre-treatment for the navy bean calibration set.

In this case, the ideal number of principal components is also 3 as the curve begins to plateau when the number of principal components becomes greater than 3. Therefore, principal components 4 through 10 capture the variance from noise in the spectra. Using 3 components captures over 95 % of the total variance.

After determining that the ideal number of principal components is 3, the loadings and scores using the PLS SIMPLS can be examined. Since the number of principal components is 3, there will be three sets of loadings and three score values for each sample. Figure 5-6 shows the three loadings which represent the weights corresponding to the wavelengths used to determine the component scores.

The loadings represent the weights of each wavelength which are then multiplied by the spectrum absorbance to give the component scores. Therefore, the highest values determine the wavelengths with the most importance. In the first loading the greatest weights occur between the wavelengths of 1900 and 2200 nm. This was the area that had the greatest difference between the protein- and starch-enriched fractions. The second loading attributes the greatest weight around the O-H peaks found between 1900 and 2000 nm. In this area there are two different O-H peaks, one associated with amines and the second associated with polymeric O-H bonds. The third loading weight has a much smaller scale and explains the least amount of variance between the samples.

The loadings are used to create the principal component scores. The scores can accurately group samples that are similar. In Figure 5-7, the principal component scores are plotted against each other as a means to separate the samples into fractions. Figures 5-7a and 5-7b accurately show splits between the protein-enriched, starch-enriched, and original flour fractions as there are separate groupings. However, using only principal components 2 and 3 as in Figure 5-7c there are no groupings for the fractions and therefore comparing these two components does not separate the fractions. .

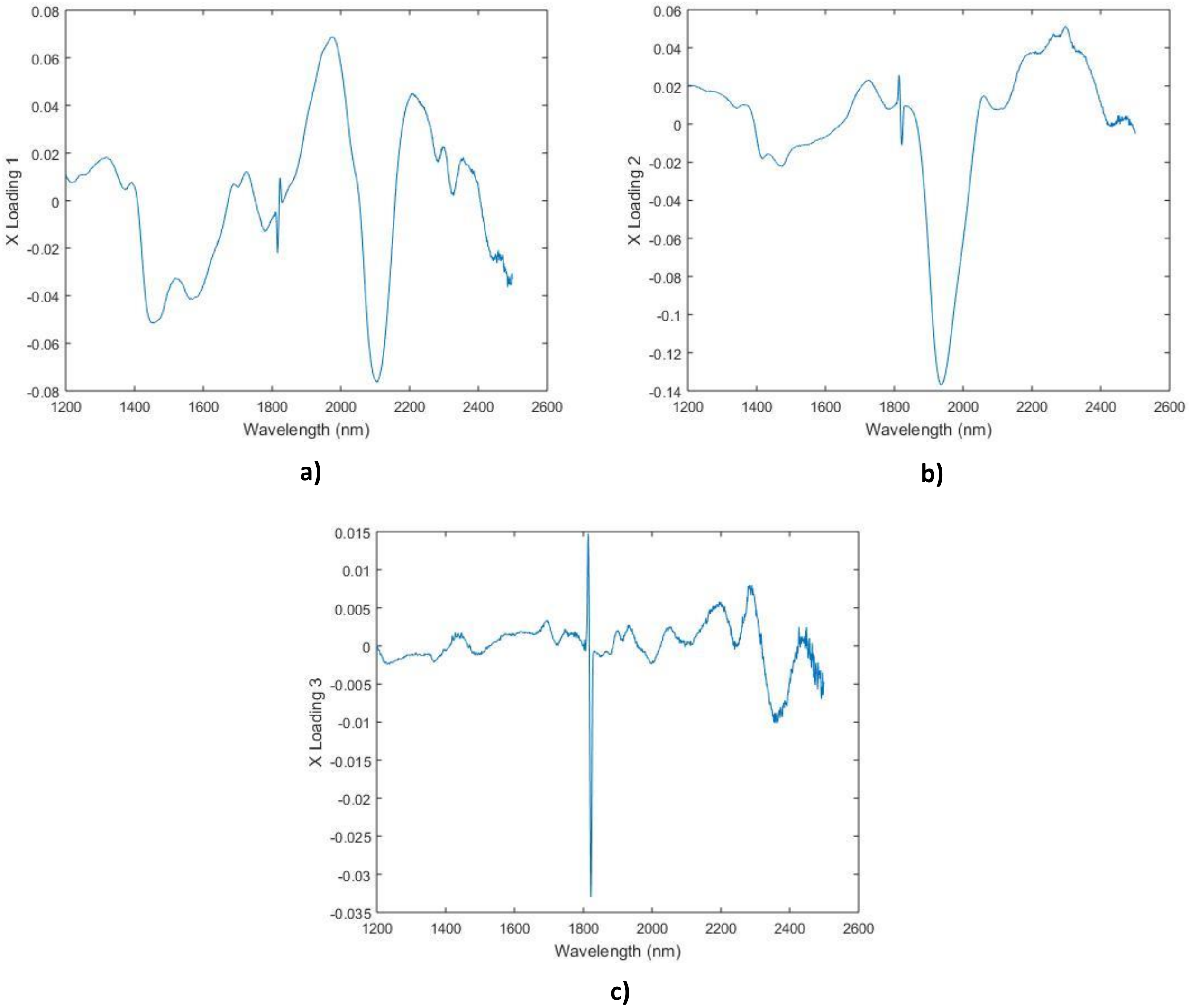


Figure 5-6: Loadings corresponding to the three principal components using PLS regression and MSC pre-treatment for the navy bean calibration set [a: first component loading; b: second component loading c: third component loading].

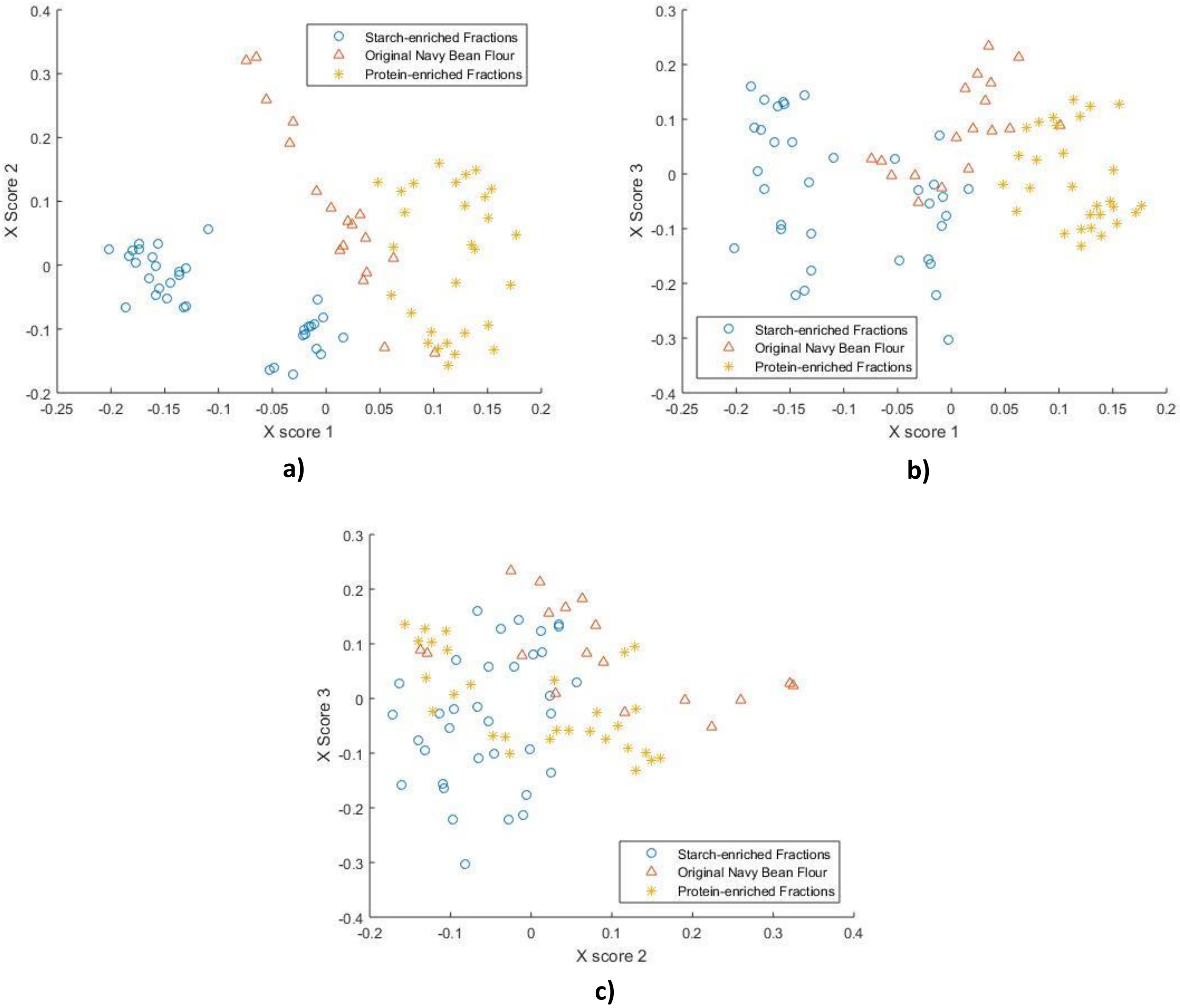


Figure 5-7: Principal component scores of protein-enriched, starch-enriched, and original flour fractions from PLS regression using MSC pre-treatment.

5.5.2 Determining Outliers

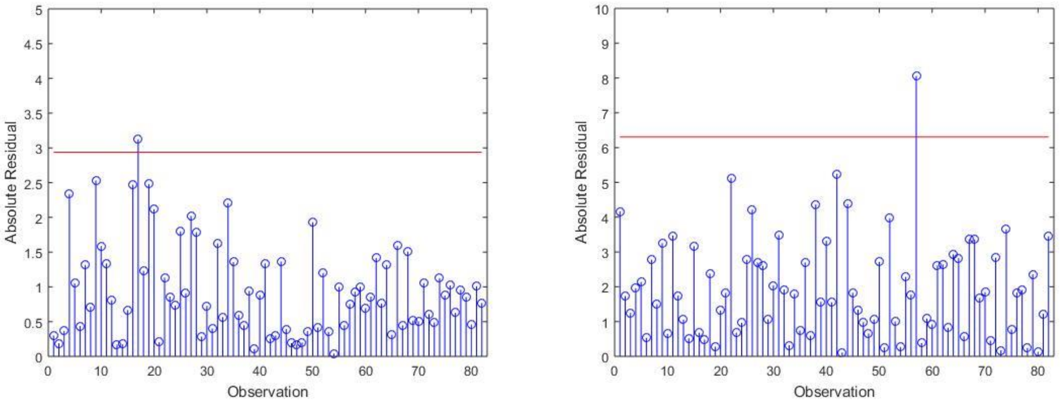
Outliers in the calibration set can severely affect the PLS model. Therefore, outliers should be determined and removed from the calibration set. An outlier is determined from the residual values which are the differences between the predicted and measured values. When a sample residual is greater than the sum of the average residual and three times the standard deviation between residuals, that sample is considered an outlier (Burns, 2008).

Figure 5-8 shows the absolute residuals for protein and starch content of each sample with a line demonstrating the threshold for outliers. There is one outlier for both protein content and starch content samples but the outliers are not the same sample. When the outliers were removed from the model, it had no effect on the final outcome. Therefore, the outliers were left in the calibration set.

Table 5-2 compares the root mean squared error of calibration and cross validation for protein content and starch content. Since the residuals for protein content on average are lower than the average residuals for starch content, it is expected that the model to predict protein content will be more accurate than starch content. The model for protein content is expected to be better for validation as the RMSECV is lower.

Table 5-2: RMSEC and RMSECV for the PLS regression model for protein and starch content.

Variable	Protein Content	Starch Content
RMSEC	1.1517	2.4365
RMSECV	1.1695	2.6022



a) b)

Figure 5-8: Absolute residuals for protein and starch content for the determination of outliers [a: protein content residuals; b: starch content residuals].

5.5.3 PLS Model Results

The PLS regression model is used to predict the protein and starch content of all the samples. There were 82 samples used in the calibration set and 20 samples used to validate the model. Figures 5-9 and 5-10 compare the model's predicted protein and starch content against the measured protein and starch content using wet analysis techniques, respectively.

The models were evaluated using the R^2 values. These values give an estimate of how close the data is to the regressed line with a value of 1 being a perfect model. Another way to determine the accuracy of the model is to compare the root mean squared errors of calibration and validation. These values give an estimate of the difference between the predicted and measured values. Table 5-3 outlines the results of the PLS regression model.

Table 5-3: PLS regression model results for protein and starch content from the NIR spectra of navy bean flour samples.

Variable	**Protein Content**	**Starch Content**
R^2 (cal)	0.973	0.940
R^2 (cal)	0.965	0.912
RMSEC	1.1517	2.4365
RMSEV	1.6826	5.9732

The protein content estimations using the PLS model are better than the starch estimates. This is seen by the higher R^2 values and lower RMSE values for both calibration and validation. However, even the starch content can be very accurately predicted given that the R^2 values are still above 0.9.

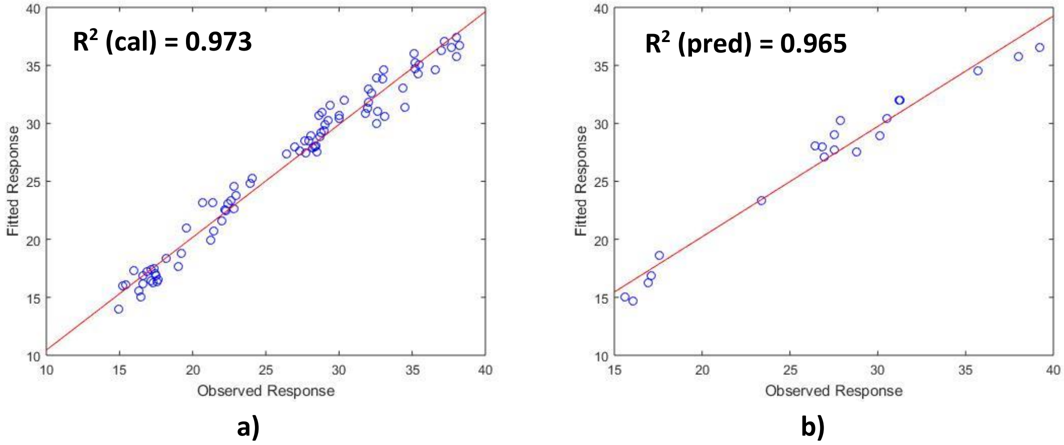


Figure 5-9: PLS Regression model for protein content of navy bean flour samples using MSC pre-treatment [a: calibration model, b: validation model].

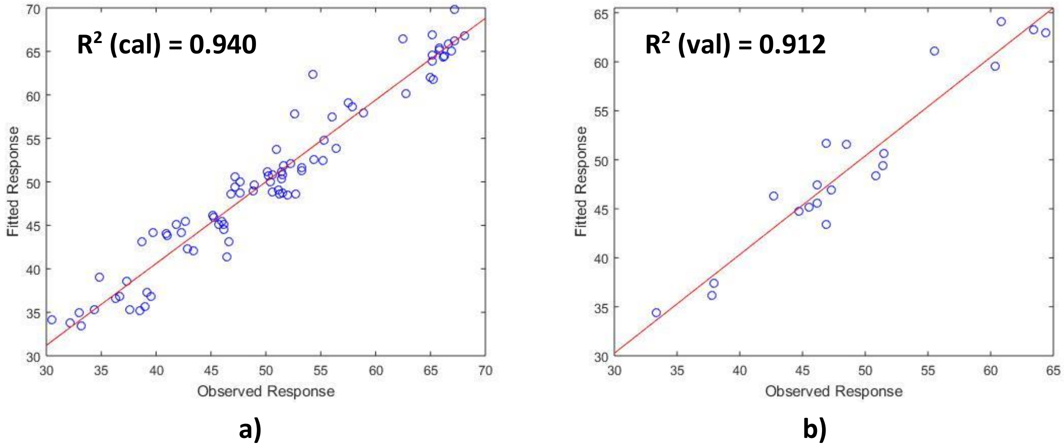


Figure 5-10: PLS Regression model for starch content of navy bean flour samples using MSC pre-treatment [a: calibration model; b: validation model].

5.6 Model Comparisons

The purpose of using multiple pre-treatments and two multivariate techniques was to determine which combination yielded the best model. Tables 5-4 and 5-5 give a summary of the results from the different model for protein and starch samples, respectively. The bolded columns represent the overall best models for each of the protein and starch samples. The best models for protein samples involve the first derivative combined with SNV and MSC pre-treatment. However, the best model for starch samples involve using just SNV or MSC pre-treatment.

The MSC pre-treated data with PLS regression as the multivariate method was chosen as the best overall method even though it was a little weaker in determining protein content because it was the best model for determining starch content. All the models very accurately predict protein content so the best overall model was the one that had the best prediction results for starch content.

Using PCR to build models performed worse than PLS regression models in terms of starch content. Also, PCR models usually required more principal components than the PLS models. This is because determining the principal components has no dependence on the measured protein and starch content values. The worst models for both protein and starch content contain no MSC or SNV pre-treatment. The first derivative and raw data created the weakest model when no further pre-treatment was added.

All models developed for protein content were stronger than then the same model's ability for determining starch content. This could be a result of the Kjeldahl protein content measurements which may be more accurate and contain less variability in repeated runs. The average standard deviation for repeated runs for the Kjeldahl protein determinations was almost half the value of the average standard deviation found in repeated DNS starch determination experiments (Appendix C).

Table 5-4: Protein content model results of navy bean flour samples for comparison of various multivariate methods and pre-treatment approaches.

	PLS						PCR					
	RAW	**MSC**	**SNV**	**1D**	**MSC+1D**	**SNV+1D**	**RAW**	**MSC**	**SNV**	**1D**	**MSC+1D**	**SNV+1D**
RMSEC	1.1128	1.1517	1.1516	1.1518	**1.0308**	**1.0301**	1.1346	1.2166	1.231	1.1519	**1.0641**	**1.0594**
RMSECV	1.1166	1.1695	1.189	1.2206	**1.0507**	**1.0463**	1.2092	1.279	1.2698	1.2062	**1.1028**	**1.0934**
RMSEV	2.0329	1.6826	1.6839	2.134	**1.4451**	**1.4396**	2.1318	1.1224	1.2404	2.1189	**1.3147**	**1.3353**
R^2 (cal)	0.9746	0.9728	0.972	0.9728	**0.9782**	**0.9782**	0.9736	0.9696	0.9649	0.9728	**0.9768**	**0.977**
R^2 (val)	0.9577	0.965	0.9649	0.9555	**0.9699**	**0.97**	0.9556	0.9766	0.9742	0.9559	**0.9726**	**0.972**
E (cal)	0.9139	0.9435	1.0549	0.9168	**0.785**	**0.7855**	0.9366	0.9559	1.0464	0.925	**0.8015**	**0.7981**
E (val)	1.1284	1.0543	1.6839	1.1495	**1.0385**	**1.0343**	1.1710	0.8496	0.8004	1.166	**0.9977**	**1.0004**
PCs	4	3	3	4	**4**	**4**	4	4	4	4	**5**	**5**

Table 5-5: Starch content model results of navy bean flour samples for comparison of various multivariate methods and pre-treatment approaches.

	PLS						PCR					
	RAW	**MSC**	**SNV**	**1D**	**MSC+1D**	**SNV+1D**	**RAW**	**MSC**	**SNV**	**1D**	**MSC+1D**	**SNV+1D**
RMSEC	2.8654	**2.4365**	**2.4366**	2.5781	2.2655	2.2671	2.9403	3.2011	2.4843	2.7911	2.3973	2.3955
RMSECV	2.8972	**2.6022**	**2.5791**	2.7209	2.579	2.5519	2.9480	2.6197	2.3718	2.9249	2.6137	2.5891
RMSEV	8.8172	**5.9732**	**5.9659**	6.9755	7.1788	7.1155	9.4003	10.0004	9.6488	7.565	6.7231	6.8056
R^2 (cal)	0.9173	**0.9402**	**0.9402**	0.933	0.9483	0.9482	0.9129	0.8968	0.898	0.9215	0.9421	0.942
R^2 (val)	0.8708	**0.9124**	**0.9125**	0.8977	0.8948	0.8957	0.8622	0.8534	0.8585	0.8891	0.9014	0.9002
E (cal)	2.2727	**1.4458**	**1.9678**	2.0792	1.9265	1.9272	2.3349	2.5278	2.4814	2.2265	1.9947	1.9954
E (val)	2.1007	**1.8709**	**1.8698**	1.9035	1.9705	1.9626	2.1787	2.2835	2.2496	1.9954	1.8203	1.8841
PCs	4	**3**	**3**	4	4	4	4	4	4	4	5	5

Chapter 6

Fluorescence Results

6.1 Data Pre-treatment

6.1.1 Scattering Effects

The same samples and fractions as described in Section 5.1 are used for this Chapter. Fluorescence data has two types of scattering: Raman and Rayleigh. Both forms originate between the molecules in the sample and the incident light. Rayleigh scattering is perfectly elastic and occurs due to molecules oscillating at a multiple of the incident light frequency. First order Rayleigh scatter would have the same frequency as the incident light and second order would have double the wavelength.

Usually only first order Rayleigh scattering is treated in fluorescence data. The common method for correcting Rayleigh scattering is to insert missing values for the width of the Rayleigh peak (Andersen 2005). However, it is difficult to determine the exact width of the Rayleigh peak, so trial and error is an option for determining this value. The peak width of Rayleigh scattering was varied from 5 to 20 nm to determine the best model. The values of emission below excitation wavelengths were set to zero as no fluorophores should emit at wavelengths lower than excitation. These variables tend to slow convergence and lead to non-ideal results (Andersen 2005).

Raman scattering is non-elastic and occurs when the emitted light has less energy than the absorbed light. It affects the entire excitation-emission matrix by the same factor; therefore, no pre-treatment method will be used to remove the effect in this study.

6.2 Multivariate Methods

6.2.1 Parallel Factor (PARAFAC) Analysis

PARAFAC analysis is used to decompose multi-way data arrays. It is like PCA and treats the different modes (variables) in the same manner (Eigenvector Research Inc. 2006). There is only one solution to a PARAFAC model and it is unique to scaling and permutation. Equation 1 shows the decomposition of the array X used in PARAFAC analysis (Murphy 2013).

$$X_{ijk} = \sum_{f=1}^{F} A_{if} B_{jf} C_{kf} + e_{if} \quad (1)$$

The nodes are equated to i, j, and k. For fluorescence excitation-emission matrices, the nodes are sample, emission, and excitation respectively. The total number of components in the PARAFAC analysis is F. PARAFAC components will estimate signals from the fluorophores if the data is approximately tri-linear (Andersen & Bro 2003). Therefore, the array X is decomposed for each variable which takes on the form of the matrices A, B, and C.

The PARAFAC model is created with some assumptions about the fluorescence data (Murphy 2013). The first assumption is that no two fluorophores will have the same fluorescence intensities or identical spectra. The next assumption is that the number of components outlines that variation between the modes in the dataset. Lastly, the total signal is a combination of the linear superposition of the fixed number of components. A change in concentration of a fluorophore will therefore only change the magnitude of the contribution and not the shape of the peak.

According to Bro & Vidal (2011), PARAFAC analysis models depend on the following factors which are all dependent on each other and cannot be chosen sequentially:

1) Wavelengths to include
2) How to handle Rayleigh scattering
3) Number of components to use
4) Samples to exclude (outliers).

The wavelengths included in the analysis here include an emission between 300 and 600 nm (increasing by 1 nm intervals) and an excitation between 250 and 380 nm (increasing by 10 nm intervals). The Rayleigh scattering will be eliminated by replacing the scatter with missing values. The width of the scatter was varied to find the best model. The number of components used was determined using core consistency (Section 5.2.2). Outliers were determined using the samples' sum squared errors with the Hotelling T^2 values. Outliers will have high residuals and/or T^2 values compared to the rest of the samples. Another method to determine outliers is to compare the score values for the principal components. An outlier would have a score value that is very different compared to all the other sample's scores.

6.2.2 Core Consistency

Core consistency is a method designed to assess the appropriateness of the PARAFAC model (Bro & Kiers 2003). The PARAFAC model is a special case of the Tucker3 model (discussed in Bro *et al.* 1998) and the Tucker3 core is used to assess the PARAFAC model. Core consistency is calculated using a binary array with zeroes in all places except the super diagonal (T) and a least squares fitted array G for a series of models with gradually increasing numbers of components. This is shown in (2) (Bro *et al.* 1998):

$$Core\ consistency = 100 * \left(1 - \frac{\sum_{d=1}^{F} \sum_{e=1}^{F} \sum_{f=1}^{F} \left(G_{def} - T_{def}\right)^2}{F}\right) \quad (2)$$

If the superdiagonal values of array G are close to one and off the superdiagonal close to 0, the model is not overfitting. A core consistency of above 90% represents very trilinear data and the number of components is acceptable. If the core consistency drops below 50%, the model is problematic and contains both trilinear and non-trilinear variation (Bro *et al.* 1998).

6.2.3 N-PLS Regression

N-PLS is a multiway calibration method for an array X of any order. It produces score vectors that have maximum covariance with the unexplained part of the dependent variable (Bro 1996). It follows the same general principal of PLS: to describe the covariance between the dependent and independent variables. It is developed as a PARAFAC-like model of array X that is obtained by modelling X as in Tucker3 decomposition (Favilla *et al.* 2013). Unlike conventional two-way PLS, there are only loading weights since the NPLS algorithm cannot calculate orthogonal scores (Eigenvector Research Inc. 2006).

Equations 3 to 7 outline the iterative algorithm for the NPLS model (Bro 1996). The first step, outlined in equation 3, is to calculate the matrix Z from the summation of the independent and dependent variables.

$$Z = X^T * Y \quad (3)$$

The next step of N-PLS is to determine the weight vectors w^J and w^K which are second and third order respectively. This is done using singular value decomposition (Jolliffe 2002). The weights are then used to calculate the scores, T, for the X array as shown in equation 4.

$$X_{ijk} = T_i * w_j^J * w_k^K \quad (4)$$

The scores of the X array are then used to solve for the regression coefficients (b) as shown in equation 5. This step is the same in two-way PLS.

$$b = (T^T * T)^{-1} * T^T * Y \quad (5)$$

After the regression coefficients have been determined, the residuals using the model are found for X and Y. When the residual values converge, the final model is found and the loop ends.

$$X_i = X_i - T_i * w^J * (w^K)^T \quad (6)$$

$$Y = Y - T * b \quad (7)$$

6.3 Qualitative Fluorescence Spectra Analysis

A fluorophore is a molecule whose chemical structure allows it to be excited by a photon and then emit that same photon while relaxing back to ground state. Molecules that can act as fluorophores mainly contain aromatic structures but can also contain highly conjugated double bonds (Christensen *et al.* 2006). Each fluorophore has an independent excitation and emission pair. Using several emission spectra at different excitation wavelengths, a 3-D excitation-emission matrix (EEM) can be created (Peiris *et al.* 2012; Peiris *et al.* 2013).

The emission spectra were gathered between wavelengths of 300 and 600 nm (increasing by one nm per interval) and for the excitation wavelengths between 250 and 380 nm (increasing by 10 nm intervals). The fluorescence spectra will be used to create a model to determine protein and starch content for navy bean flour fractions from the electrostatic separation (Section 3.3). Protein content can be determined using the fluorescence that can be attributed to tryptophan. This peak resides at a excitation/emission pair of approximately 295/330 nm (Sharma & Kalonia 2003).

Tyrosine also fluoresces, however its intensity is far lower than that of tryptophan (Calvet *et al.* 2012). Therefore, if tryptophan is present in the protein, it is not possible to see the contribution of tyrosine. Figure 6-1 shows the excitation-emission spectra using front-face fluorescence for whey protein isolate. This figure clearly shows the location of the tryptophan peak. It is also worth noting there is a second peak for which at this point the chemical attribution is unknown.

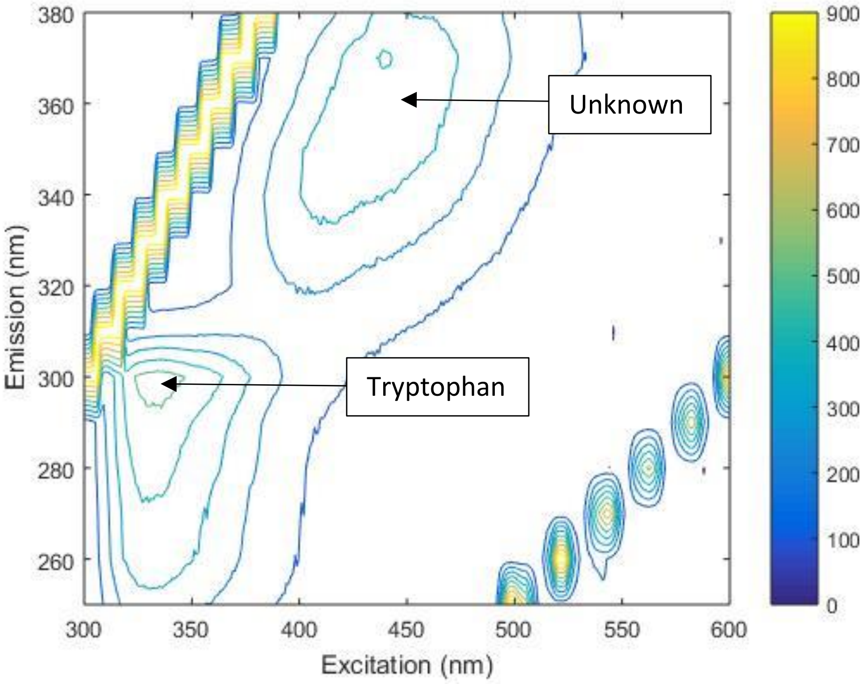


Figure 6-1: Excitation-emission matrix of whey protein isolate acquired using front-face fluorescence.

Sugars do not possess the chemical structures that produces any fluorescence. Figure 6-2 provides the excitation-emission matrix of potato starch using front-face fluorescence.

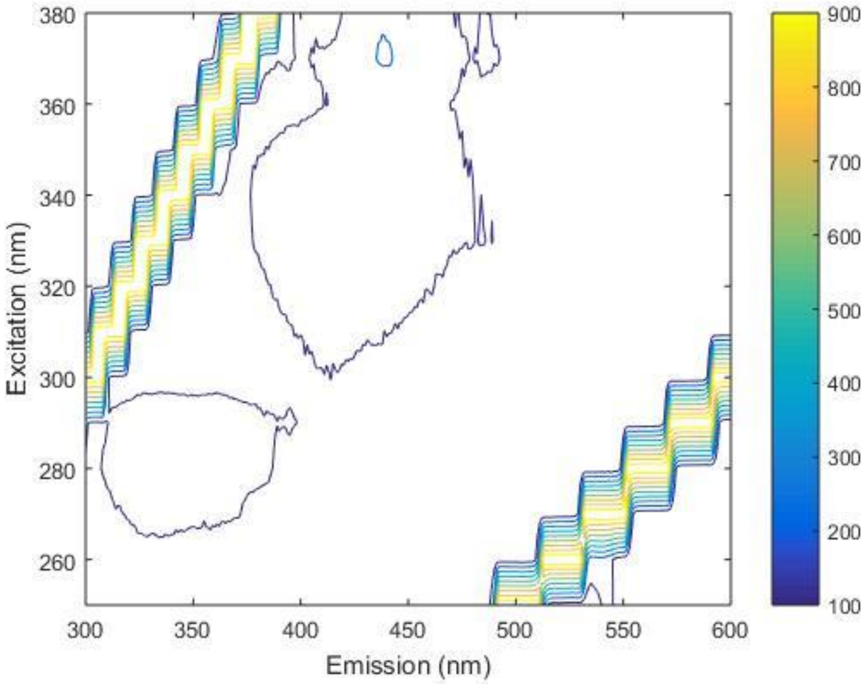


Figure 6-2: Excitation-emission matrix of potato starch acquired using front-face fluorescence.

As shown in Figure 6-3, there are two small peaks located at the tryptophan and unknown peak. However, the intensities of the peaks are much smaller than what is found in the whey protein isolate. There are no other additional peaks shown in the excitation-emission matrix of starch. The model quantifying starch content from the fluorescence data should be based on a lack of protein content or smaller fluorophore peaks.

The navy bean flour in this study contains both protein and starch in different proportions and therefore the resulting excitation-emission matrix should be a combination of what is seen Figures 6-1 and 6-2. Figure 6-3 is the front faced fluorescence excitation-emission matrix acquired for raw navy bean flour.

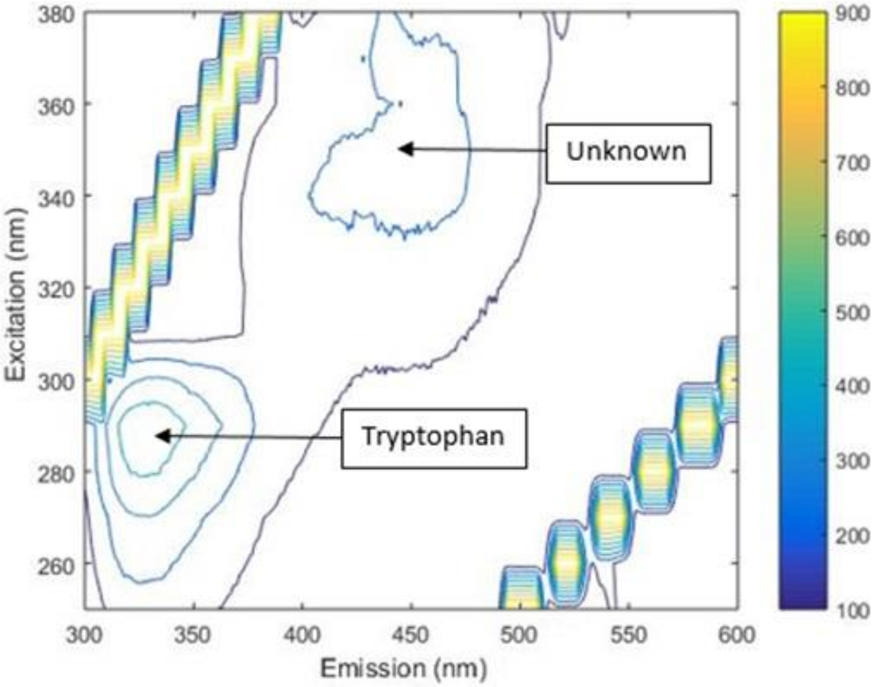


Figure 6-3: Excitation-emission matrix for raw navy bean flour without any pre-treatment.

The tryptophan and unknown peaks are identified in Figure 6-3. The intensities of these peaks are similar in value to the whey protein isolate sample reported in Figure 6-1. Pre-treatment of the raw fluorescence data is necessary to remove the Rayleigh scattering data and as well as any emission wavelength which is lower than the excitation wavelength. Figure 6-4 shows the excitation-emission matrix of the raw navy bean flour after the data has been pre-treated. The pre-treatment of the fluorescence data should improve the speed of obtaining the PARAFAC and n-PLS models as well as the model's accuracy (Andersen 2005).

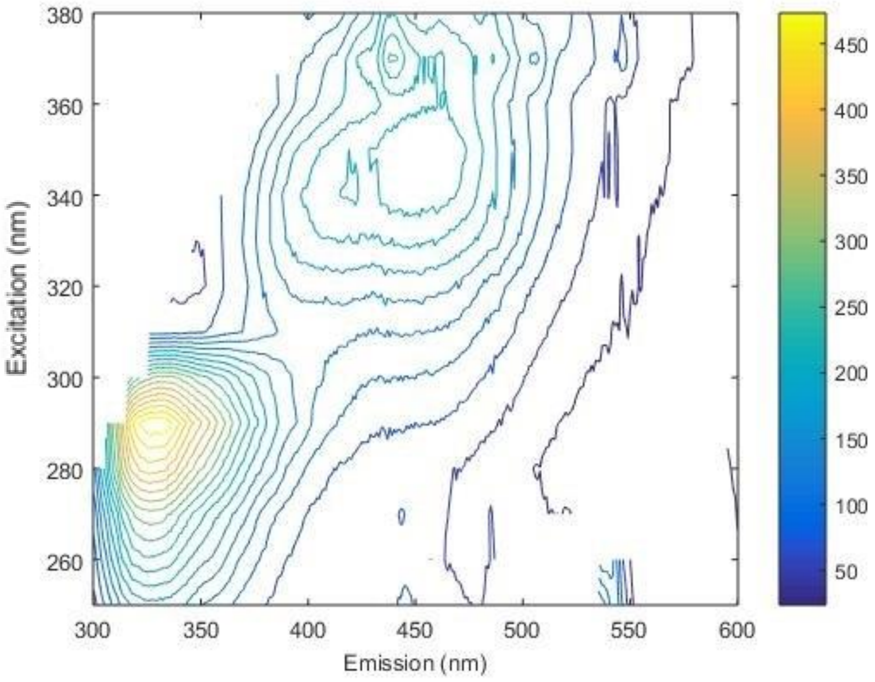
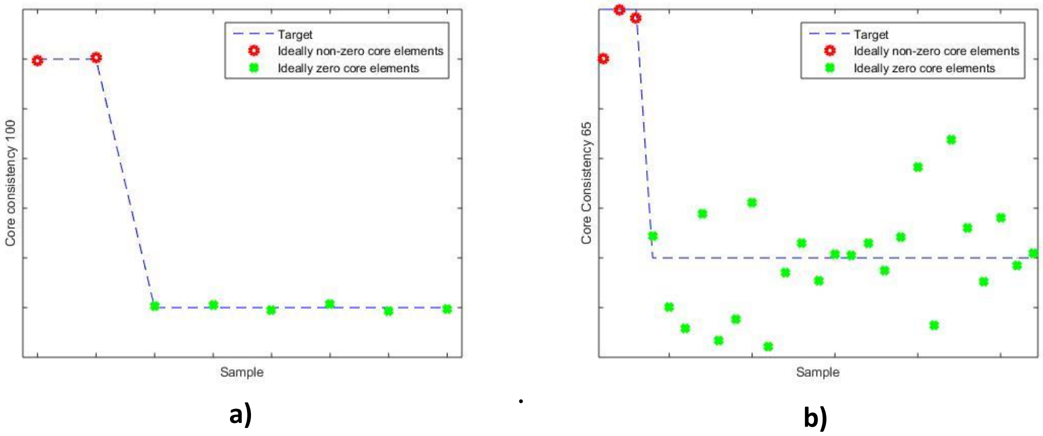


Figure 6-4: Excitation-emission matrix for navy bean flour with pre-treatment including removal of the Rayleigh scattering and wavelengths of emission less than excitation wavelength.

6.4 PARAFAC Model

6.4.1 Number of Principal Components

The first step of PARAFAC modelling is to determine the number of principal components ideal for the dataset. This is done using core consistency. Figure 6-5 shows the core consistency results for 2 and 3 components.



a)

b)

Figure 6-5: Core consistency of front-faced fluorescence spectra of navy bean flour fractions [a: 2 components, b: 3 components].

As seen in Figure 6-5, the ideal number of components is 2. When the number of components is increased to 3, the ideally zero core elements vary from the target values which causes the core consistency value to decrease from 100. The core consistency of 2 components is 100 while with 3 components this value drops to 65.

When using 2 components the PARAFAC model explains over 99% of the variation shown in the fluorescence matrix. Figure 6-6 depicts the amount of variance explained in each of the 2 principal components. PARAFAC modelling is unlike PCA because the principal components are not orthogonal and therefore principal components have the ability for components to repeat the same feature which is better described using only 1 component. The two-component model shows a slightly higher amount of variance explained in the first component.

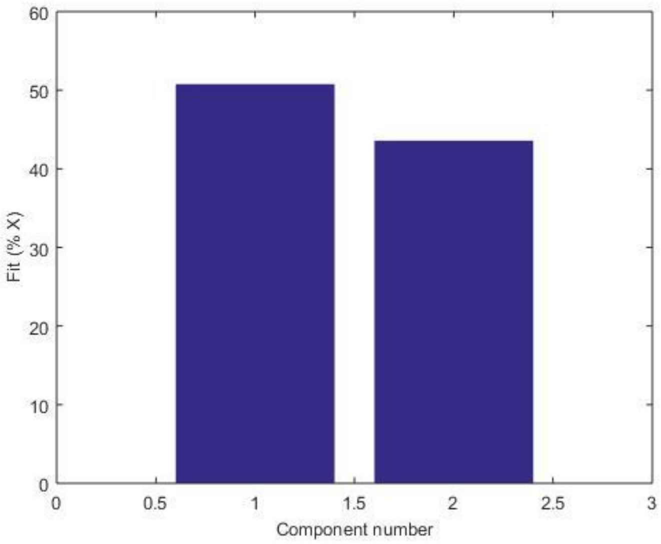


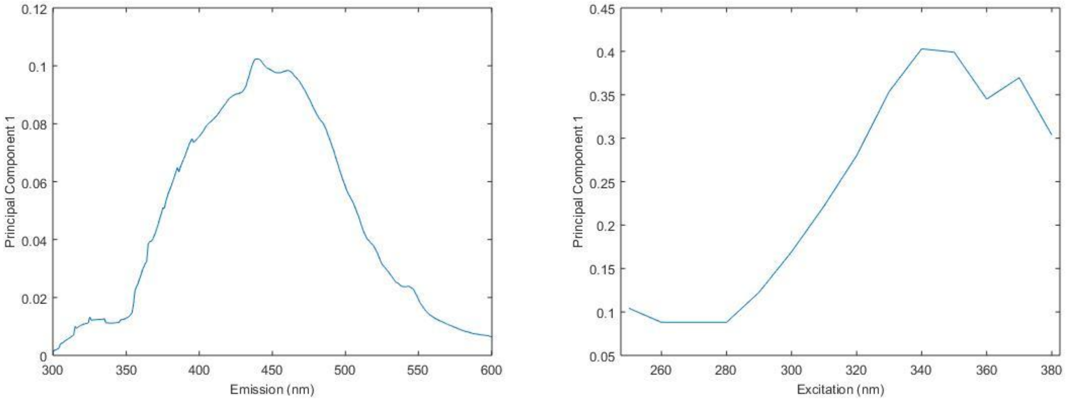
Figure 6-6: Percent variance explained for the fluorescence matrix for both components in the PARAFAC model.

6.4.2 Principal Component Loadings and Scores

The principal component loadings for PARAFAC models are attributed to individual fluorophores. Each principal component contains one peak, both from an excitation and emission standpoint. With two principal components, two fluorophores for the navy bean flour samples are expected. Two peaks were identified from the raw navy bean flour excitation-emission matrix (found in Figure 6-3).

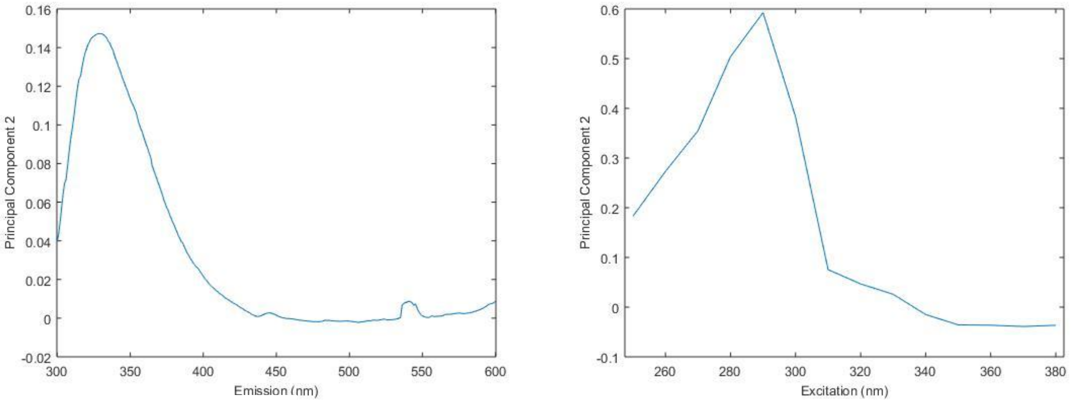
The first principal component represents a fluorophore to which assignment of chemical identity is uncertain and is shown in Figure 6-7. This peak is relatively wide as it has a emission base of approximately 200 nm. It also has an irregular shape, as the curve itself is not smooth in the emission curve. The excitation curve is not expected to be smooth as the data points are 10 nm apart. It has a shape that is similar to humic acid (Lee *et al.* 2015; Yan *et al.* 2013), which is ruled out because it is not found in navy bean flour.

The second principal component, shown in Figure 6-8, ist attributed to tryptophan as its emission maximum occurs at 330 nm with an excitation at 290 nm (Sharma & Kalonia 2003). This peak has a higher intensity than the unknown fluorophore and a smooth narrow shape. There is a second peak in this principal component which could be the result of the influence of the unknown peak as it is also seen in the first prinicpal component.



a) b)

Figure 6-7: First principal component of the PARAFAC model representing the unknown fluorophore [a: emission curve, b: excitation curve].



a) b)

Figure 6-8: Second principal component of the PARAFAC model representing tryptophan [a: emission curve, b: excitation curve].

The component scores of each sample can be used to group the 82 navy bean fraction samples that were analyzed for starch and protein content. The samples were grouped into protein-enriched, starch-enriched, and raw navy bean flour fractions. Figure 6-9 shows the principal component scores plotted against each other for the three fractions.

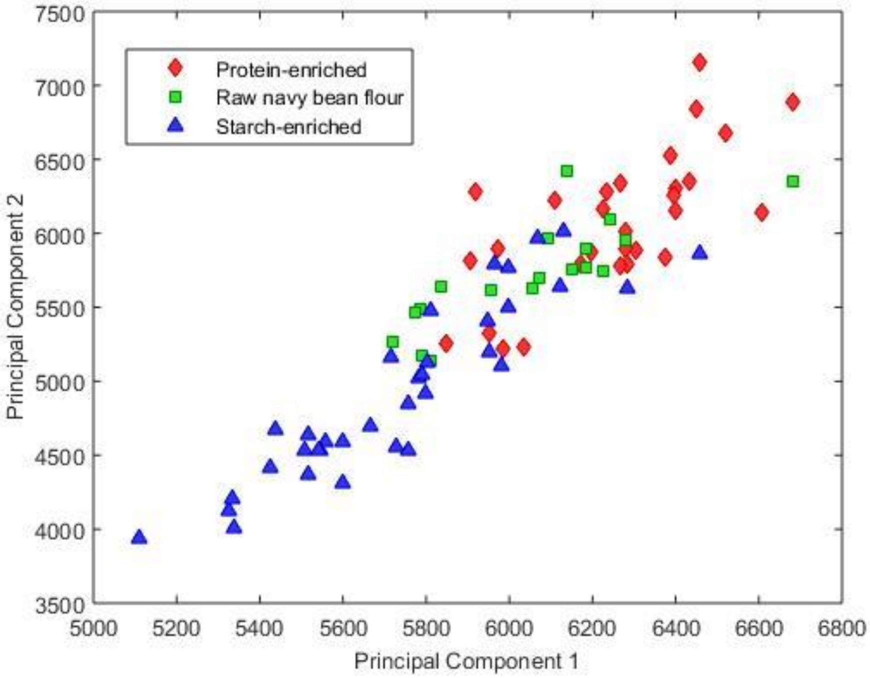


Figure 6-9: Principal component scores for protein-enriched, starch-enriched, and raw navy bean flour fractions.

The protein-enriched samples tend to have higher scores while the starch-enriched fractions tend to have lower scores for both principal components. There is not a clear separation between the fractions but the overall trend is apparent. The protein enriched samples possess higher score values of both components. This means that both components explain the protein content of the samples.

6.4.3 Residuals and Outliers

Sample outliers are determined using the plot of Hotelling T^2 and sum squared errors. Samples that have much higher T^2 or sum squared error values are outliers. The plot of Hotelling T^2 and sum squared errors for the samples is shown in Figure 6-10.

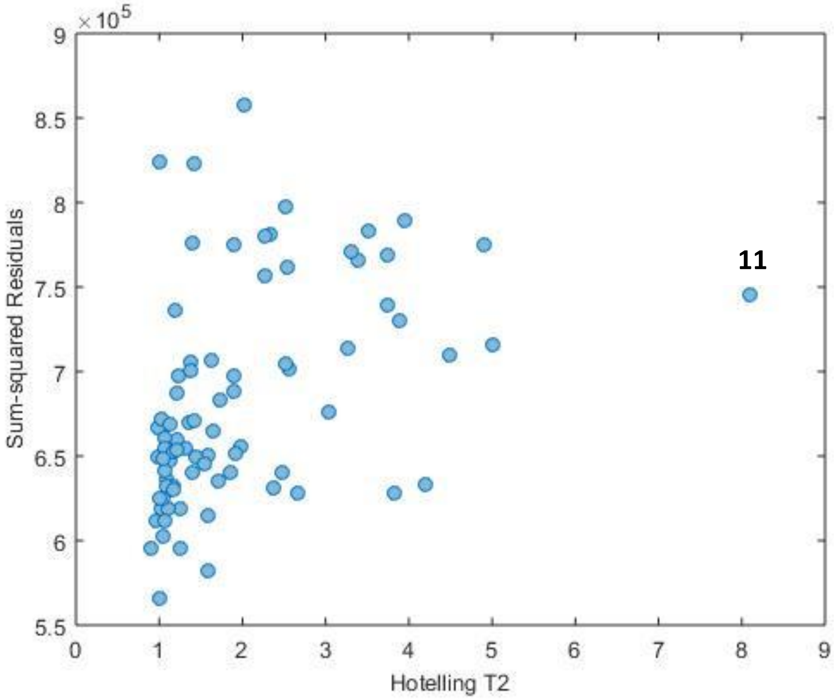


Figure 6-10: Influence plot comparing Hotelling T^2 and sum squared error values for each navy bean flour sample.

As shown in Figure 6-10 there is only one sample with a much higher Hotelling T^2 value than the rest of the samples (sample 11). Since the sample's sum squared error is similar to all the other samples and the Hotelling T^2 is not an order of magnitude greater, sample 11 (protein-enriched sample) is not deemed to be an outlier and is included in the PARAFAC model. Figure 6-11 compares the sum squared error for the three fractions.

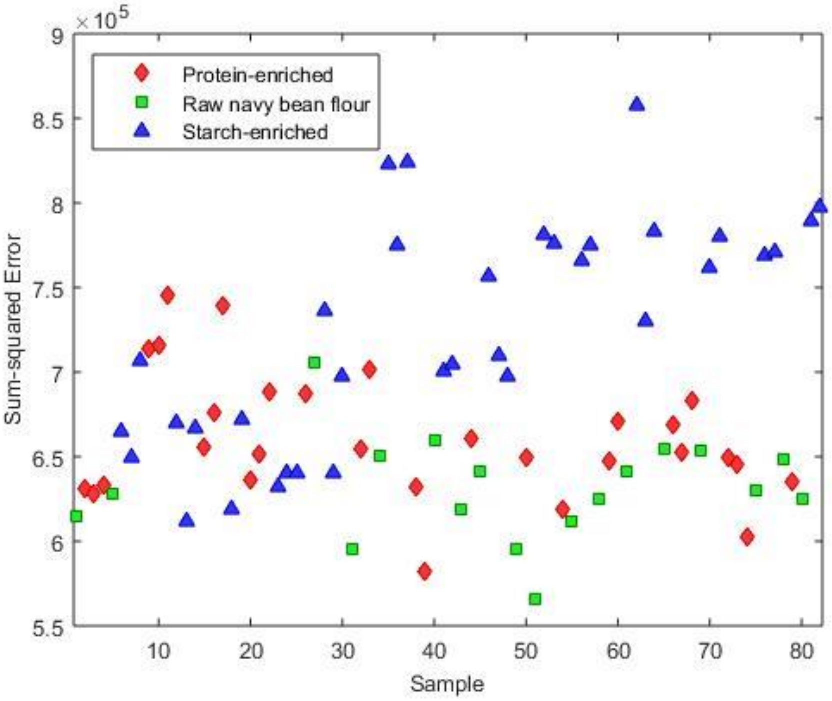


Figure 6-11: Sum-squared errors of the three fractions from the PARAFAC model.

The starch-enriched samples on average have a higher sum squared error than the raw navy bean flour and protein-enriched fractions. This may be due to starch not containing any major fluorophores. The excitation-emission matrix of potato starch (approximately 99% pure) did possess small peaks for both tryptophan and the unknown fluorophore. This may explain why the starch-enriched samples tend to have higher sum squared errors in the PARAFAC model. Lastly, no samples are observed as outliers when comparing all the sample's sum squared errors.

6.5 N-PLS Model

6.5.1 Principal Components

Determining the ideal number of principal components is the first step of NPLS modelling. The goal is to use the most principal components without overtraining the model to the calibration set. An overtrained model will perform poorly when predicting samples outside the calibration set. Cross validation is a powerful tool to determine the number of principal components. Figure 6-12 shows the root mean squared error of calibration and cross validation for both protein and starch content.

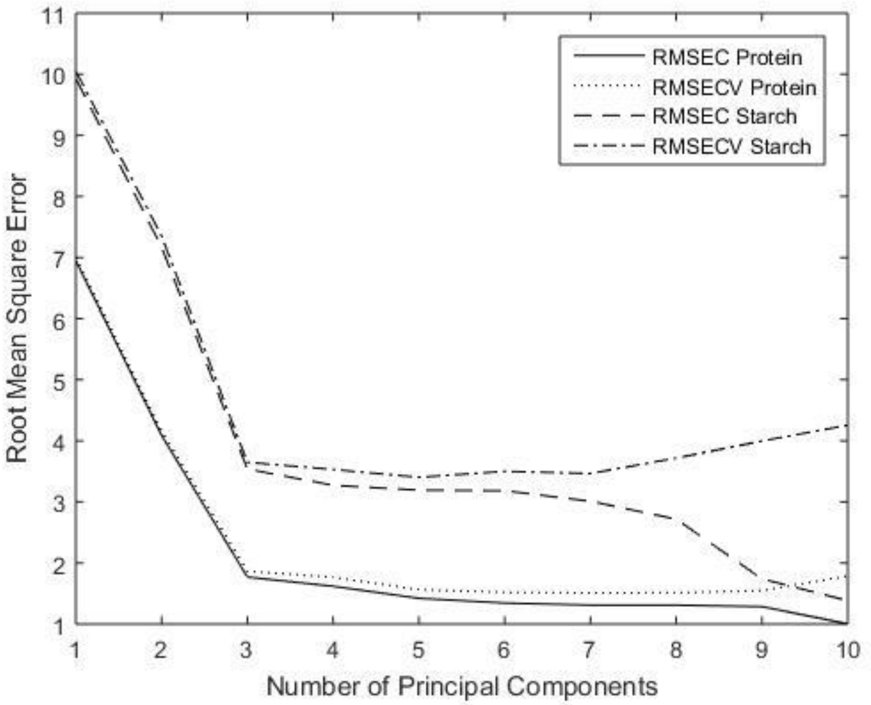


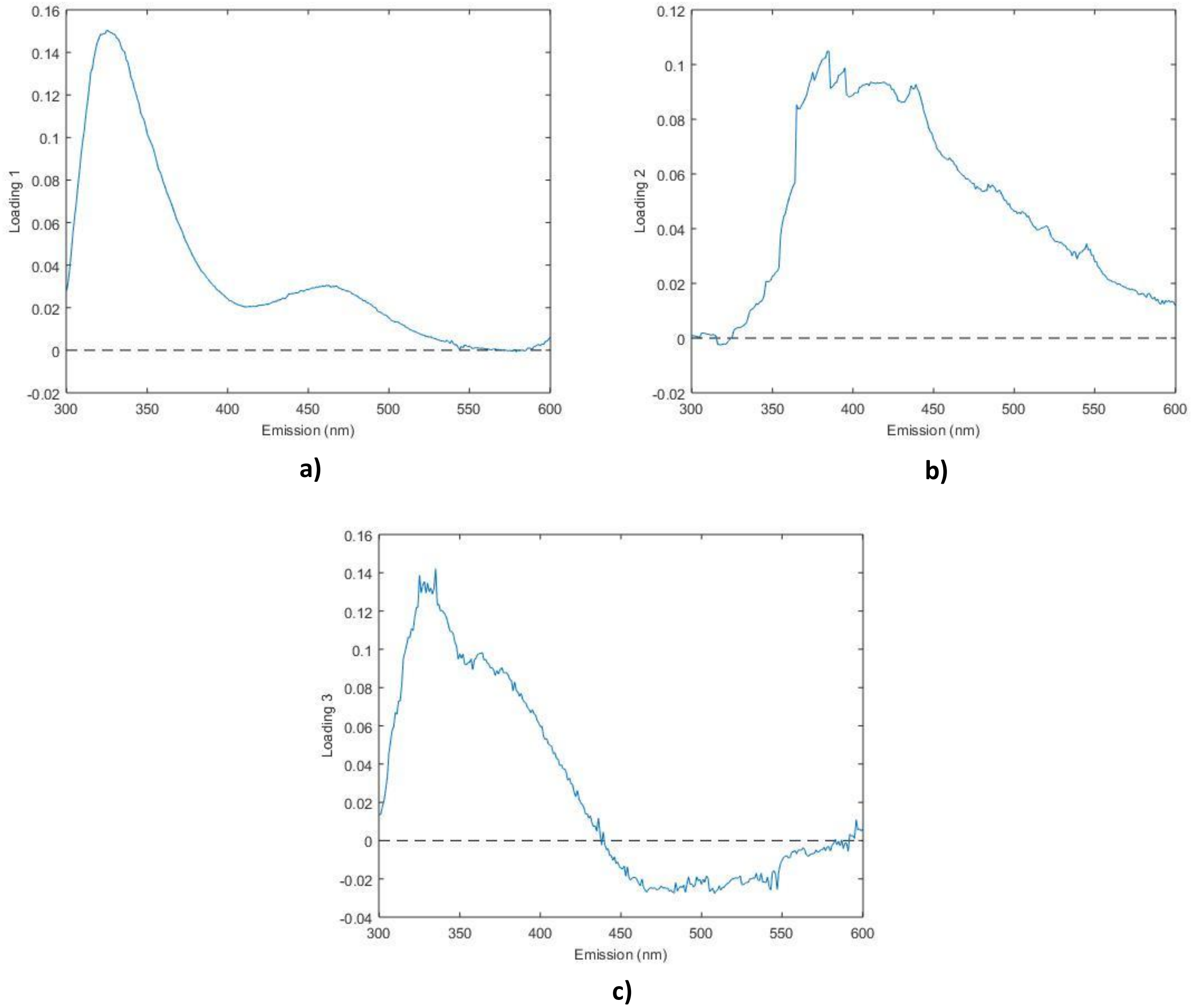
Figure 6-12: Root mean square error of calibration and cross validation for protein and starch content to determine the number of principal components.

The ideal number of principal components occurs when the root mean squared error of cross validation (RMSECV) plateaus or increases as the number of principal components increase. According to Figure 6-12, the number of principal components should be 3 which will be used in the NPLS model.

After determining that the ideal number of principal components is three, the loadings and scores can be examined. Since the number of principal components is three, there will be three sets of loadings and three score values for each sample. Since the fluorescence matrix is 3D (contains both excitation and emission wavelengths) there will be 3 loadings for both the emission and excitation wavelengths. Figure 6-13 shows the emission wavelength loadings while Figure 6-14 shows the loadings for the excitation wavelengths.

The loadings for emission wavelengths in the NPLS are similar to the loadings found in the PARAFAC model (Section 5.4). The first two emission loadings appear to be the same shapes as the tryptophan and unknown fluorophore peaks. The third principal component seems to be a combination of the two fluorophores with greater reliance on the tryptophan area. The tryptophan peak shape (found in the first loading) has a smoother shape than the other two principal components.

The excitation wavelength loadings consist of only 14 data points as there were only 14 excitation wavelengths analyzed. The pattern with the excitation loadings is the same as the emission loadings. The first loading once again appears to correlate with the tryptophan peak, the second loading is attributed to the unknown fluorophore, and the third a combination of the two.



a)

b)

c)

Figure 6-13: Emission wavelength loadings corresponding the three principal components using NPLS regression [a: first component loading, b: second component loading, c: third component loading].

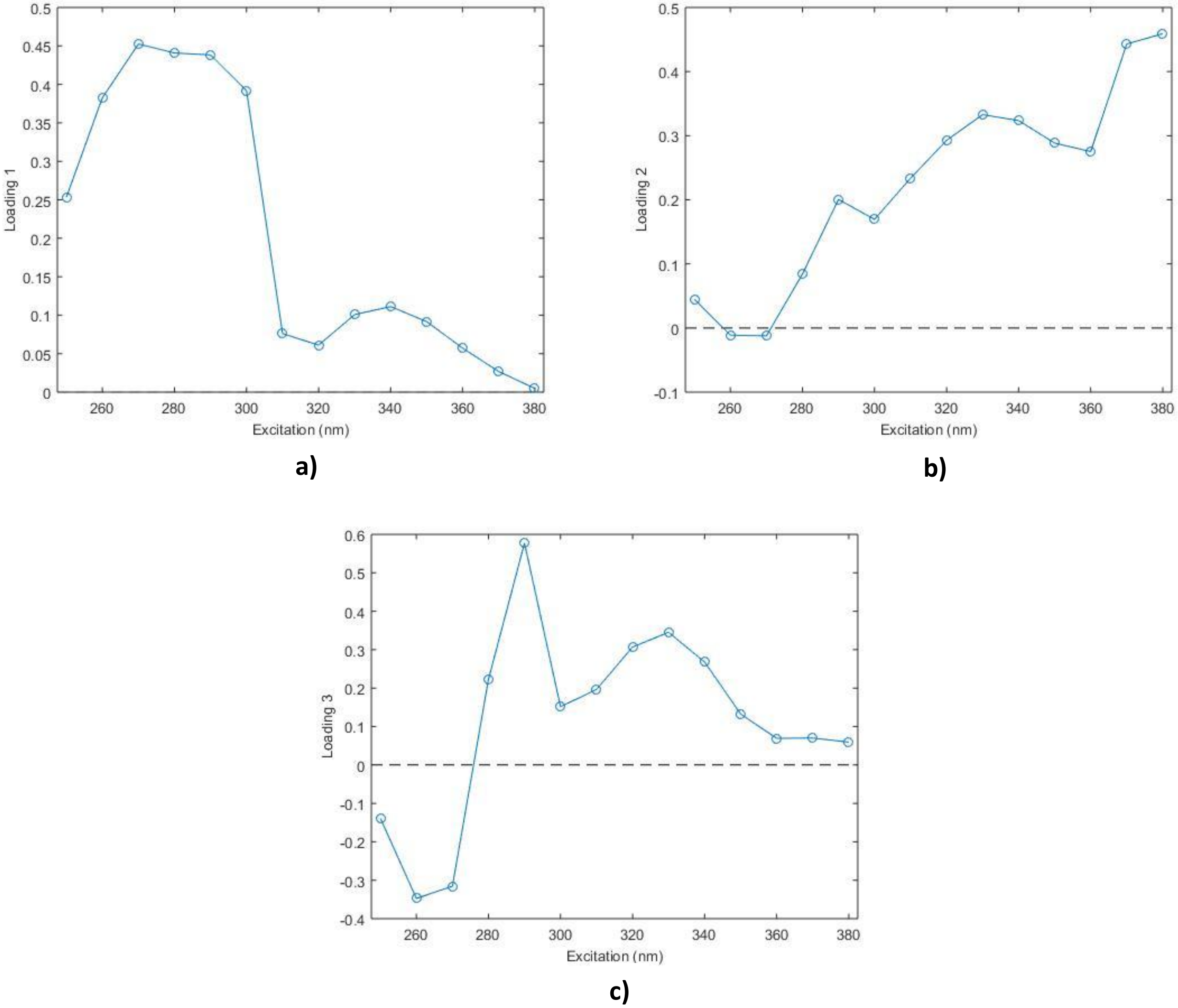


Figure 6-14: Excitation wavelength loadings corresponding the three principal components using NPLS regression [a: first component loading, b: second component loading, c: third component loading].

The combination of excitation and emission loadings are used to create the principal component scores. Figure 6-15 plots the principal component scores against each other as a means to separate the samples in sections. The groupings are split into: protein-enriched, raw navy bean flour, and starch-enriched. The first component score describes the most variance in the excitation-emission wavelengths with 53.7% and the total variance captured between the three components is over 94%.

Since the first two components capture the most variance, it is expected that comparing these two scores will yield the best comparison between the three fractions. However, there are no clear divisions between the fractions but an overall trend is examined. The protein-enriched samples tend to have higher values while the starch-enriched samples tend to have lower values for components 1 and 2. The third component score only accounts for about 4% of the total variance and therefore shows the least separation between fractions.

The principal component scores of the calibration set are used to create the final NPLS model. Since there is much less separation between fractions than what was shown in the NIR data, it is expected that the fluorescence model for determining protein and starch content should be weaker.

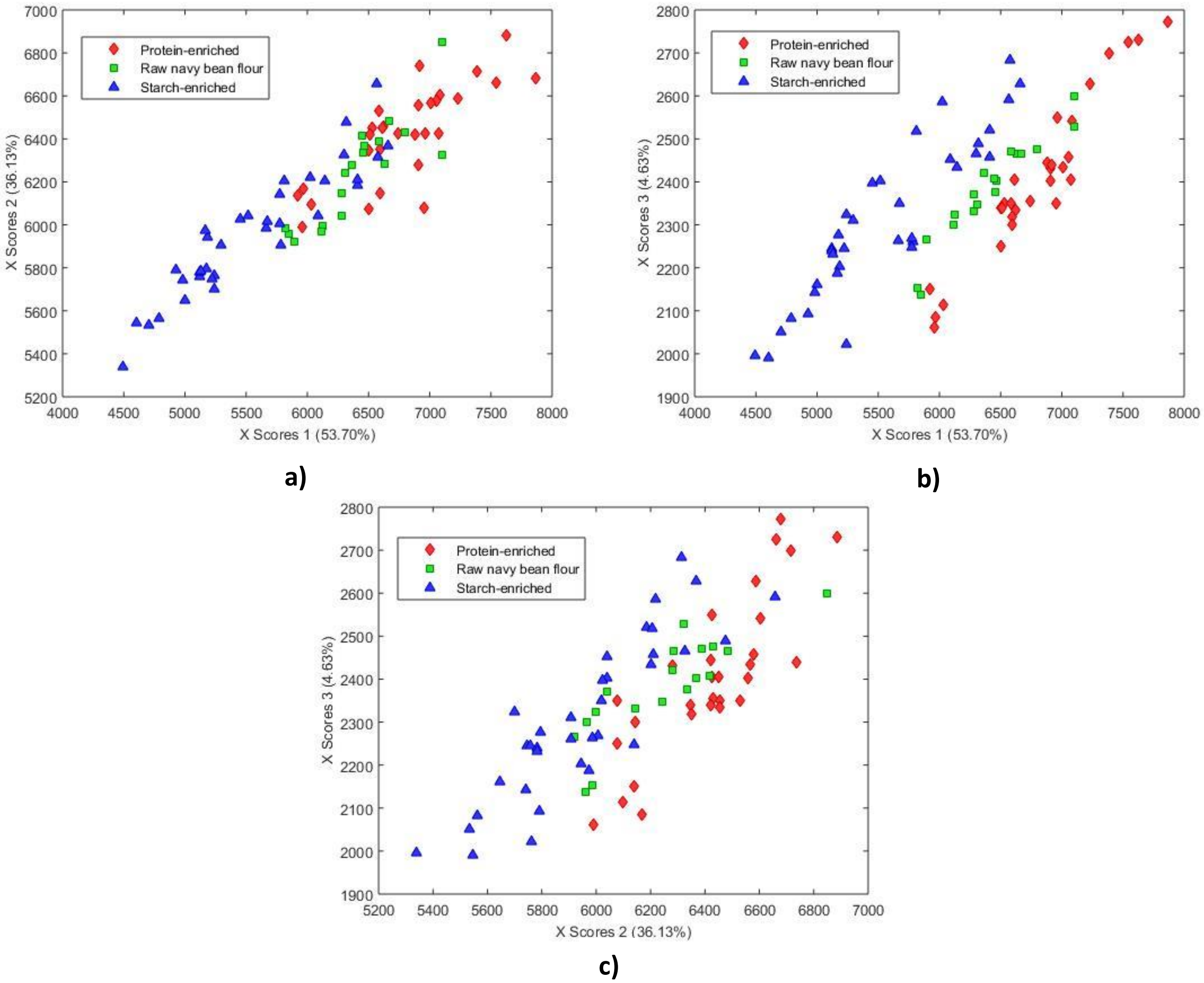


Figure 6-15: Principal component scores of protein-enriched, starch-enriched, and original flour fractions from NPLS regression [a: scores 1 and 2, b: scores 1 and 3, c: scores 2 and 3].

6.5.2 Determining Outliers

Outliers in the calibration set can severely affect the PLS model. Therefore, outliers should be determined and removed from the calibration set. An outlier is determined from the residual values which are the difference between the predicted and measured values. A sample will be deemed an outlier if the residual value is high and removing the sample significantly affects the model.

Figure 6-16 shows the residual values for protein and starch content of each sample. The sample with the highest residual value for both protein and starch content was removed from the calibration set and a new model was developed. In both cases there was not a significant change in the model and therefore the samples were included in the final calibration set.

The average residual for protein content is lower than for starch content. Table 6-1 compares the root mean squared error of calibration and cross validation for protein content and starch content. Since the root mean squared error of calibration and cross validation is lower for protein content, it is expected that the model to predict protein content will be more accurate than starch content.

Table 6-1: RMSEC and RMSECV for the NPLS regression model for protein and starch content.

Variable	Protein Content	Starch Content
RMSEC	1.7685	3.5377
RMSECV	1.8668	3.6470

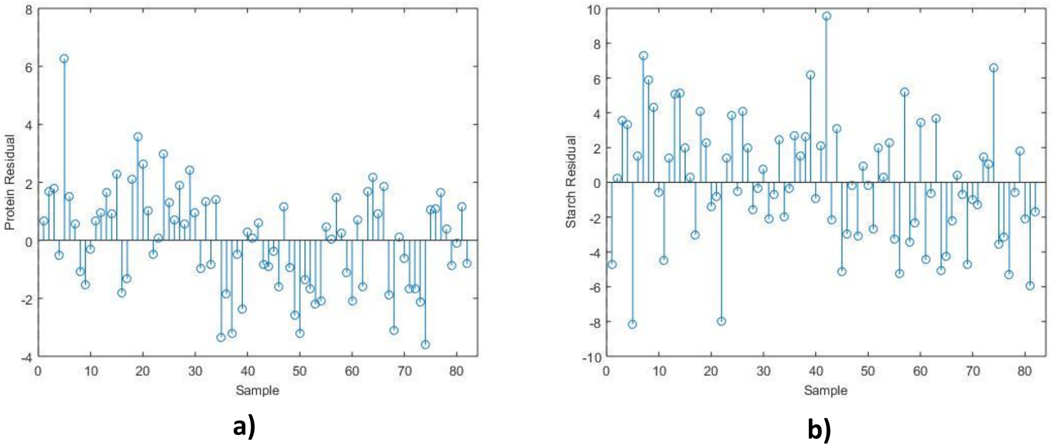


Figure 6-16: Residuals for protein and starch content for the determination of outliers [a: protein content residuals, b: starch content residuals].

6.5.3 NPLS Model Results

The NPLS regression model is used to predict the protein and starch content of all the samples. There were 82 samples used in the calibration set and 20 samples used to validate the model. Figures 6-17 and 6-18 compare the model's predicted protein and starch content against the measured protein and starch content, respectively for both the calibration and validation samples.

The models were evaluated using the R^2 values. These values give an estimate of how close the data is to the regressed line with a value of 1 being a perfect model. Another way to determine the accuracy of the model is to compare the root mean squared errors of calibration and validation. These values give an estimate of the difference between the predicted and measured values. Table 6-2 outlines the results of the PLS regression model.

Table 6-2: PLS regression model results for protein and starch content.

Variable	Protein Content	Starch Content
R^2 (cal)	0.936	0.874
R^2 (val)	0.946	0.885
RMSEC	1.7685	3.5377
RMSEV	1.9102	2.8558

The protein content estimations using the PLS model are better than the starch estimates as was the case for the NIR spectra. This is seen by the higher R^2 values and lower RMSE values for both calibration and validation. The model for starch is still a good model since the R^2 values are just under 0.9. The R^2 values for validation are higher than for calibration which means that the model is very robust.

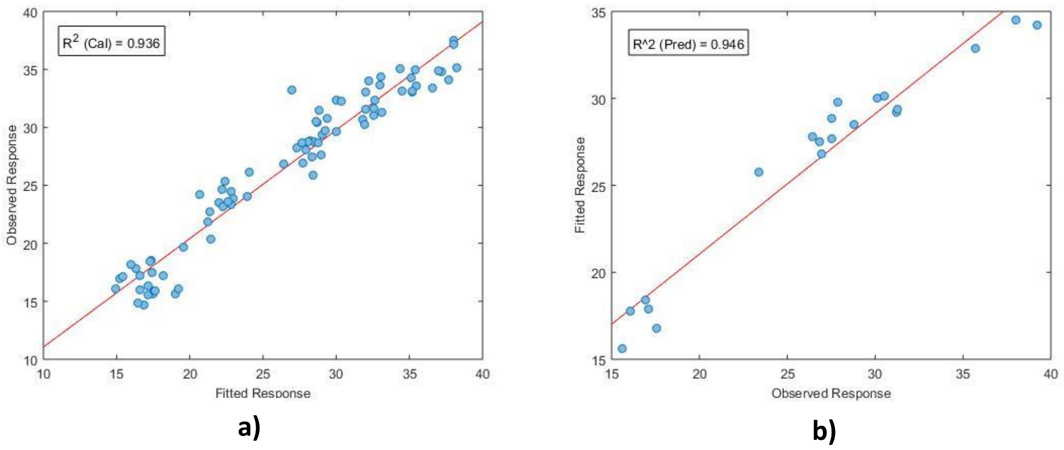


Figure 6-17: NPLS Regression model for protein content using fluorescence spectra [a: calibration model, b: validation model].

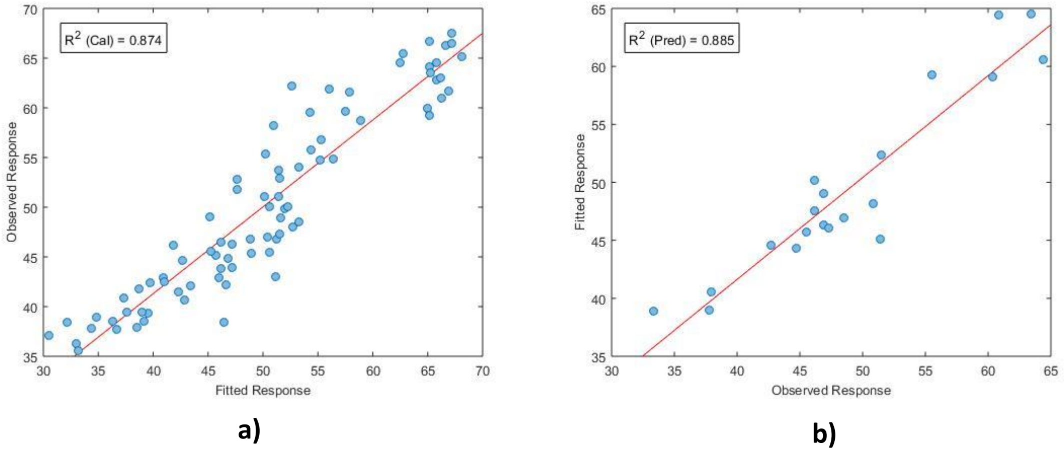


Figure 6-18: NPLS Regression model for starch content using fluorescence spectra [a: calibration model, b: validation model].

6.5.5 Starch Content Model

As discussed earlier in this chapter, starch does not contain the necessary chemical components to fluoresce; therefore, the prediction of starch content must be determined from the unknown fluorophore or could even be related to the predicted protein content. Since the scores for the starch-enriched samples had lower scores for principal component 2 (describing the unknown fluorophore), the starch content prediction is most likely a function of the predicted protein content. Figure 6-19 shows the relationship between the predicted values of protein and starch content.

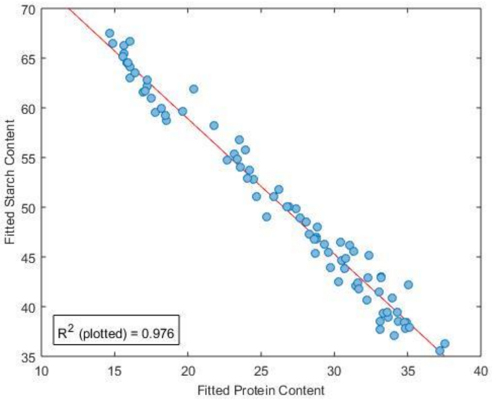


Figure 6-19: Predicted protein and starch content relationship from the NPLS regression model.

As shown in the above plot, the predicted starch content demonstrates a better relationship with the predicted protein content than the actual measured values. The R^2 value is extremely close to 1, signifying an almost perfect relationship.

Chapter 7

Data Fusion of NIR and Fluorescence Spectra

7.1 Data Pre-treatment and Fusion Setup

The same samples and fractions as described in Section 5.1 are used for this Chapter. As discussed in Section 2.5, there are three different levels of data fusion. Low level data fusion was used to combine the NIR and fluorescence spectral data. This data fusion type allows the application of specific pre-treatment strategies before combining the data (Solano *et al.* 2012). SNV pre-treatment was used for the NIR data while the fluorescence spectral data had no pre-treatment.

An important aspect of data fusion is to adjust the NIR and fluorescence data to the same scale. Therefore, each set of data was scaled to range from 0 to 1 by subtracting the minimum data point and dividing by the maximum. Lastly, it is important for each set of data to have approximately the same number of data points (Khaleghi *et al.* 2013). This helps to ensure that the model for the combined data is not weighted more heavily on one of the fused portions more than the other. The setup of the NIR and fluorescence data prior to fusion is discussed in the following sections.

7.1.1 NIR Data Setup Prior to Fusion

The number of NIR data points was not reduced; therefore, the only setup necessary was the SNV pre-treatment and scaling of the absorbance between 0 and 1. Figure 7-1 depicts the SNV pre-treated data and SNV scaled data for raw navy bean flour, protein-enriched, and starch-enriched samples.

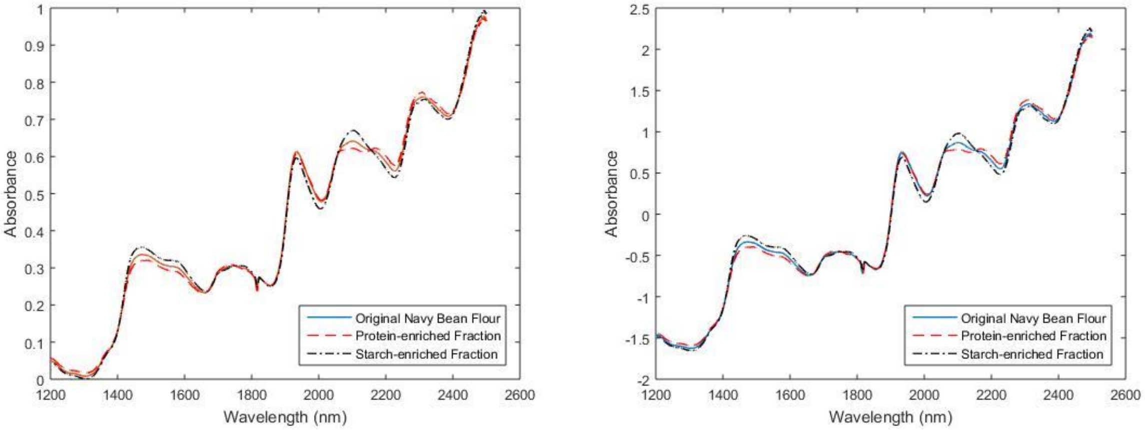


Figure 7-1: NIR spectra of navy bean flour, protein-enriched, and starch-enriched samples for data fusion [a: SNV pre-treated data, b: SNV scaled pre-treated data].

7.1.2 Fluorescence Data Setup Prior to Fusion

No individual pre-treatment was necessary for the fluorescence data; however, the number of fluorescence data points has to be approximately the same as the NIR data points. Since there are only 1301 NIR data points, approximately 75 % of the fluorescence data points have to be omitted.

As in Chapter 6, there are two fluorophores present in the excitation-emission spectra for navy bean flour (tryptophan and an unknown fluorophore). This is the area where there is the most variance between protein-enriched and starch-enriched samples and it is precisely this area that was chosen from the fluorescence spectral data. Figure 7-2 shows the approximate area (shaded region) taken from the fluorescence spectra to be used in the data fusion outlined in Table 7-1.

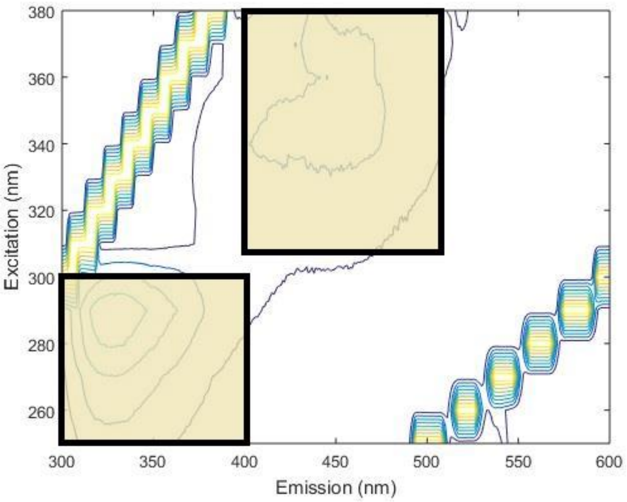


Figure 7-2: Fluorescence spectra of navy bean flour with the shaded regions used in the data fusion with the NIR spectral data.

Table 7-1: Summary of fluorescence data points used in the data fusion with the NIR spectral data.

Excitation (nm)	Emission Range (nm)	Number of data points
250	300-350	50
260	300-375	75
270	300-375	75
280	300-375	75
290	300-400	100
300	313-400	87
310	400-500	100
320	400-500	100
330	400-500	100
340	400-500	100
350	400-500	100
360	400-500	100
370	400-500	100
380	400-500	100
	Total	**1262**

Since the NIR spectral data is 2-dimensional and the fluorescence spectral data is 3-dimensional, it is necessary to unfold the fluorescence spectral data. Lastly, the fluorescence data was also scaled between 0 and 1. Figure 7-3 shows the chosen fluorescence data points, outlined in Table 7-1, unfolded and scaled for raw navy bean flour, protein-enriched, and starch-enriched samples.

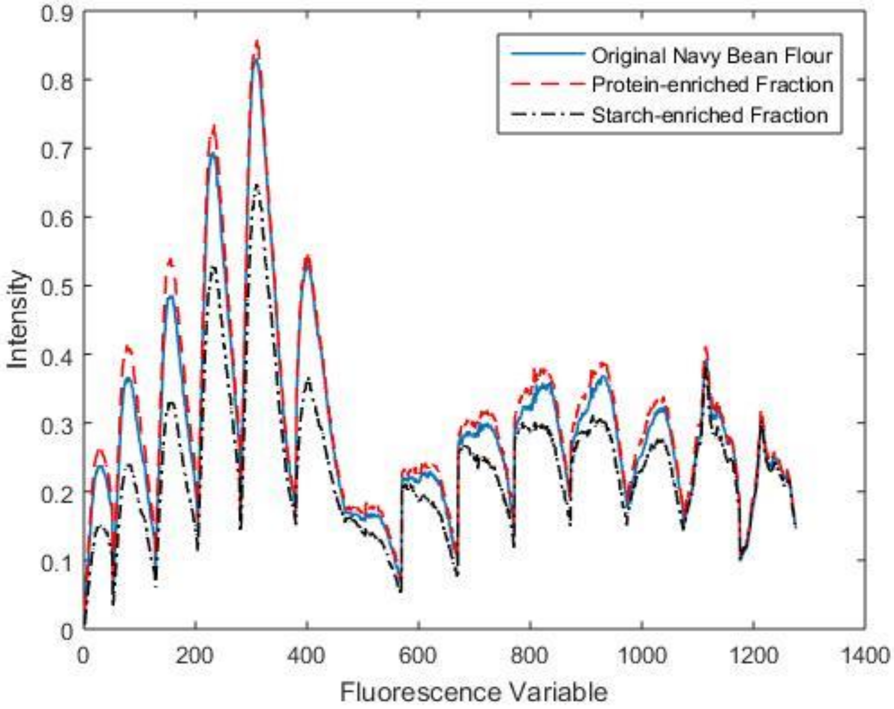


Figure 7-3: Unfolded and scaled fluorescence spectra of raw navy bean flour, protein-enriched, and starch-enriched samples for data fusion.

7.1.3 Fused Data

A summary of the pre-treatment and data reduction for the NIR and fluorescence spectra is shown in Figure 7-4. The final fused spectra have 2577 data points.

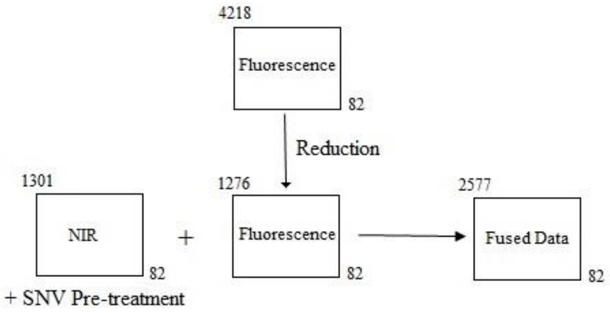


Figure 7-4: Summary of approach used for pre-treatment, data reduction, and fusion of the NIR and fluorescence spectral data.

The final fused data has approximately the same number of data points from the NIR and fluorescence spectra. Figure 7-5 represents the fused data for raw navy bean flour, protein-enriched, and starch-enriched samples. The NIR data is the first 1301 data points and the unfolded fluorescence spectra follows. The NIR and fluorescence data was conserved to better examine the correlation between the data and the model loadings.

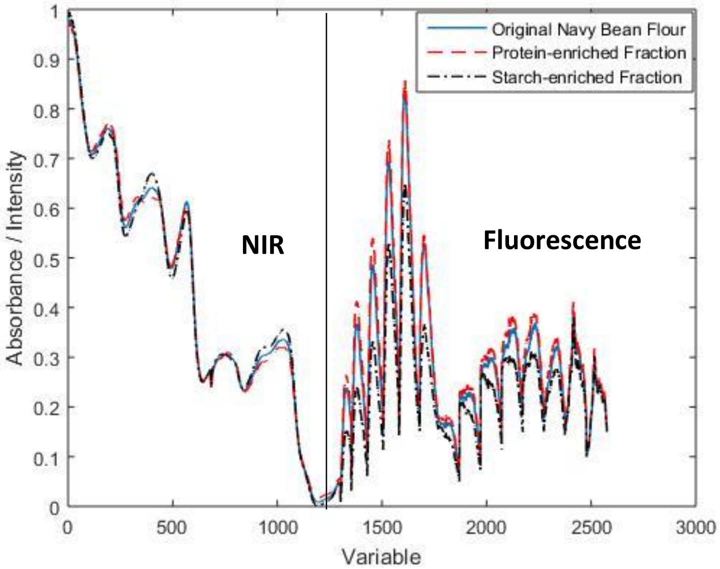


Figure 7-5: Fused NIR and fluorescence spectra for raw navy bean flour, protein-enriched, and starch-enriched samples.

7.2 NIR and Fluorescence Fused Data PLS Model

PLS regression was chosen to create the fused data's model for protein and starch content. In Chapter 5, it was determined that using PLS produced better models for starch content than PCR and the starch content model has greater room for improvement.

7.2.1 Number of Components

The ideal number of components for the model is determined from cross validation. Figure 7-6 shows the root mean squared error of calibration and validation for protein and starch content. The number of principal components was varied between 1 and 10.

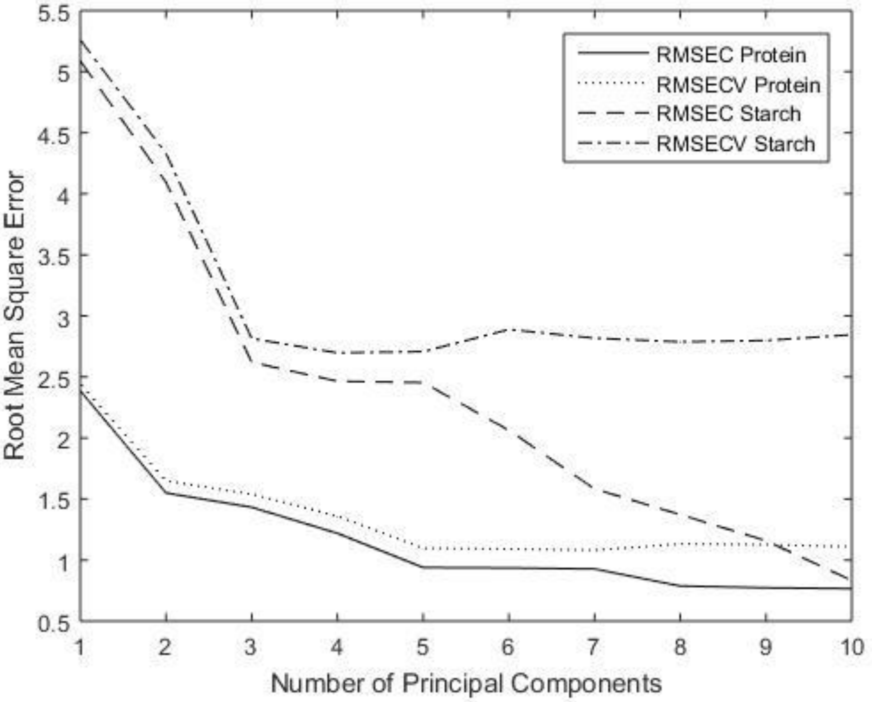


Figure 7-6: RMSECV and RMSEC for protein and starch content using 1 to 10 principal components with the fused NIR and fluorescence data.

The root mean squared error of cross validation (RMSECV) for protein content plateaus at five components. However, the RMSECV for starch content plateaus earlier at three components. In Chapter 6 it was determined that the starch content model for the fluorescence data depended only on the model for protein content and not the two fluorophores present in the navy bean flour samples. Additionally, a three component model was the best for the NIR spectra with SNV pre-treatment. Therefore, it was expected that the ideal number of principal components for starch content should not increase from three by fusing the fluorescence spectra.

The fluorescence spectra did yield a good relationship with the protein content via the tryptophan and unknown peaks. Also, the NIR spectra had peaks relating to the amide group which is characteristic of protein (Workman & Weyer 2012). This explains the larger number of principal components that are able to explain the protein content from the fused NIR and fluorescence data. Since the RMSECV for starch does not significantly increase from 3 to 5 principal components, the PLS model was developed using 5 components.

7.2.2 Principal Component Loadings and Scores

There will be five sets of loadings and score values for each sample since the number of principal components is five. Figure 7-7 shows the five loadings which represent the weights corresponding to the fused data. The five principal components explain over 95 % of the variance in the fused data matrix. Most the variance is found in the first two components as the remaining account for less than 10 % of the total variance each.

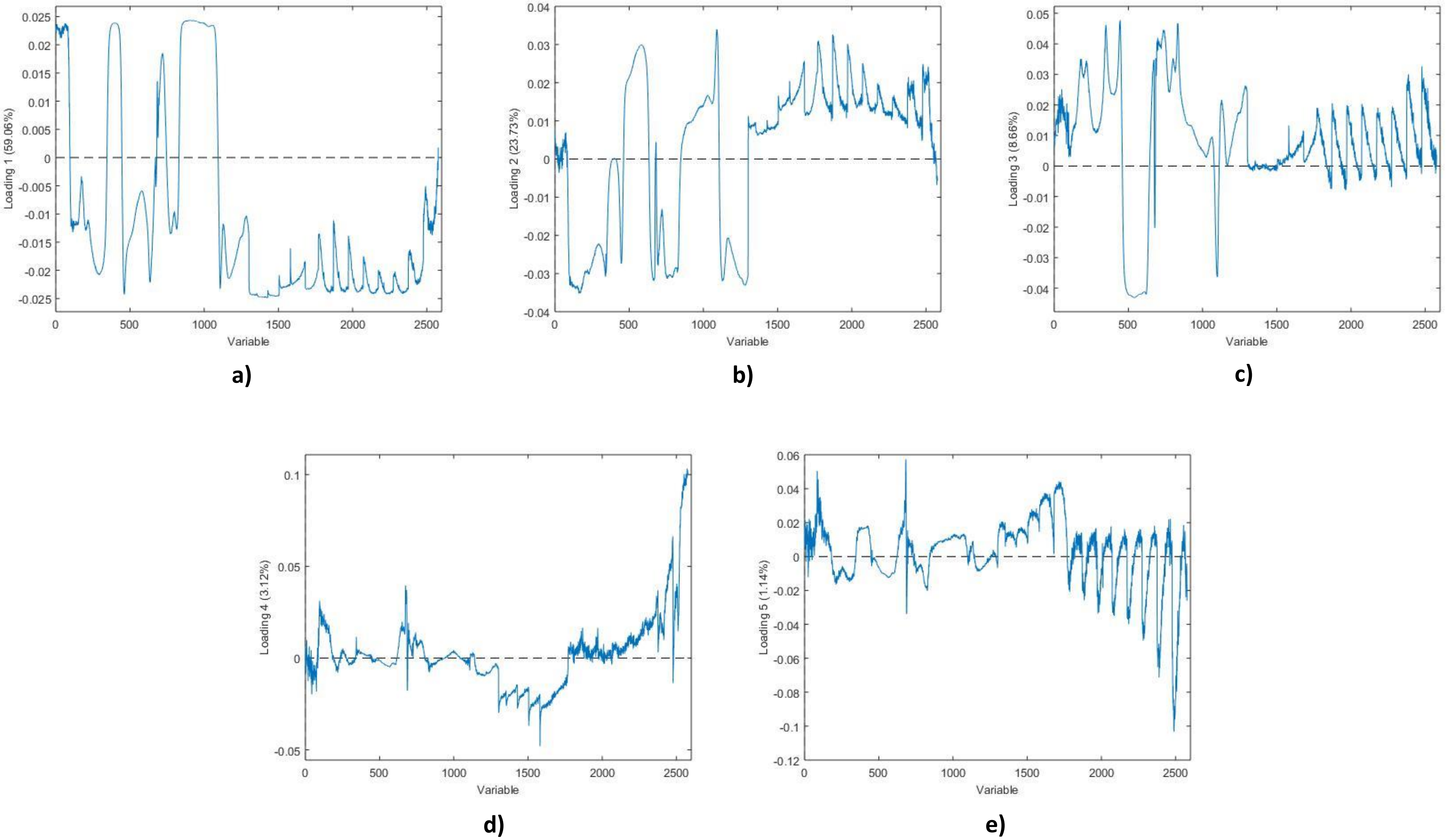
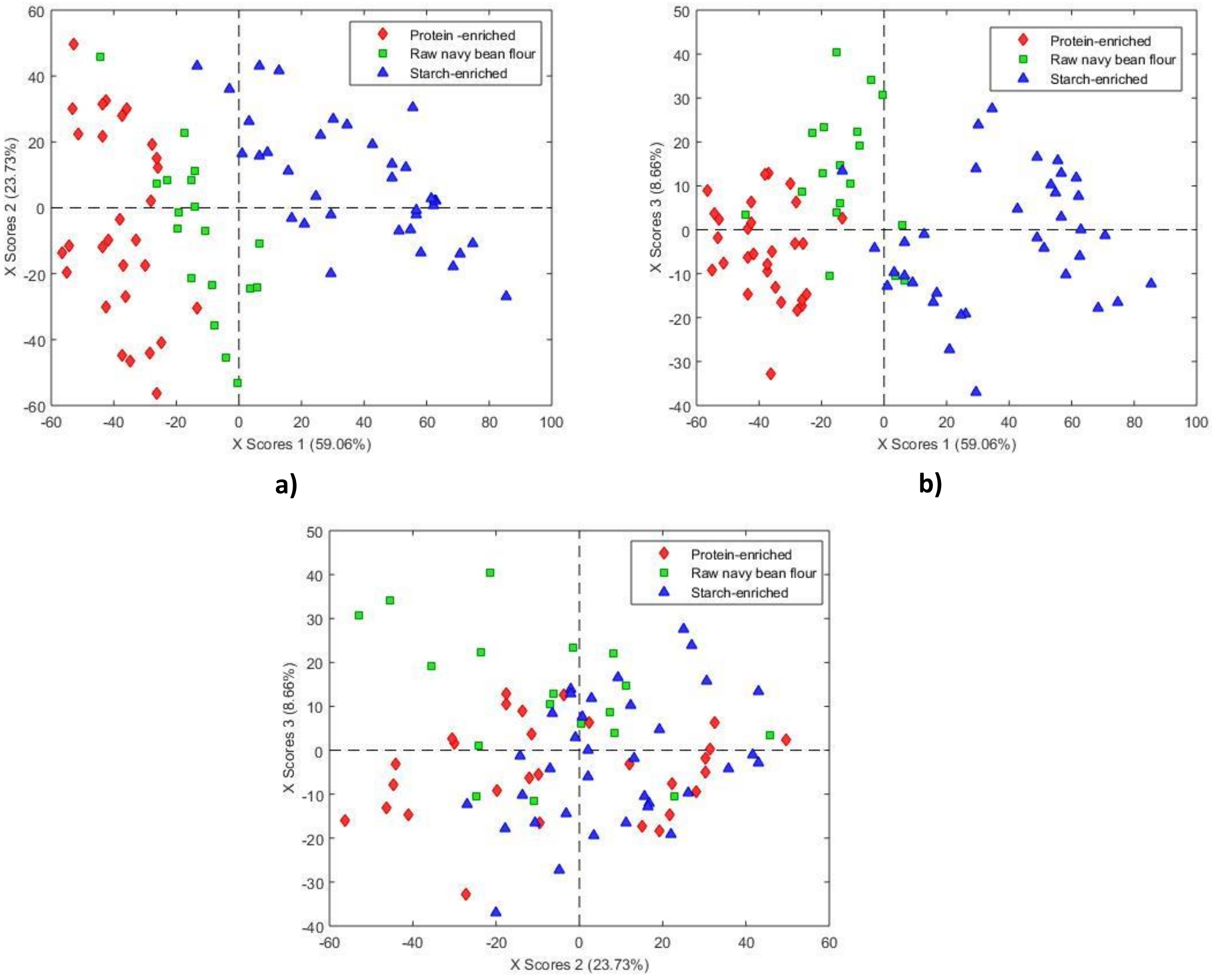


Figure 7-7: Loadings corresponding the five principal components using PLS regression on the NIR and fluorescence fused data [a: first component loading, b: second component loading, c: third component loading, d: fourth component loading, e: fifth component loading].

All principal component loadings are the resulting weights of the NIR and fluorescence data. That is each principal component loading has weights for the NIR and fluorescence data with no loading being singular to one of the fused data sets. The first two loadings show that in the NIR region of the fused data, the highest weights are assigned to the 1900 to 2200 nm wavelength range. In the fluorescence region, the shape is similar to the actual fluorescence section of the fused data. The highest weights in this section seem to be the peaks attributed to the primary fluorophores attributed to the protein.

The principal component loadings are combined with the fused data to create the principal component scores. The scores are used in PLS regression to create the final model. In Figure 7-8, the first three principal component scores are plotted against each other to separate the samples into fractions. The fractions are raw navy bean flour, protein-enriched, and starch-enriched samples.

As shown in Figures 7-8a and 7-8b, there are clear groupings of the three fractions. Using any combination of scores that does not include the first component results in no grouping or pattern found between the fractions. This is shown in Figure 7-8c as the second and third components are plotted against each other. The score plots that include the fourth and fifth component are not shown because there are also no groupings between the fractions and are like Figure 7-8c which is expected because both of these principal components represent less than 5 % of the total variance in the fused data matrix.



a)

b)

c)

Figure 7-8: Principal component scores of protein-enriched, starch-enriched, and original flour fractions for fused NIR and fluorescence data from PLS regression [a: first/second component scores, b: first/third component scores, c: second/third component scores].

7.2.3 Residuals and Outliers

As discussed in previous Chapter 5, outliers in the calibration set can severely affect the PLS model and should be determined and removed from the calibration set. A sample will be deemed an outlier if the residual value is high and removing the sample significantly affects the model.

Figure 7-9 shows the residual values for protein and starch content of each sample. The sample with the highest residual value for both protein and starch content was removed from the calibration set and a new model was developed. In both cases there was not a significant change in the model and therefore the samples were included in the final calibration set.

The average residual for protein content is lower than for starch content which means that the model for protein content should be better. Table 7-2 compares the root mean squared error of calibration and cross validation for protein content and starch content. The root mean squared values are also lower for protein content, also suggesting that the protein content model will have a stronger relationship with the fused data. Lastly, the average absolute residual for protein content is extremely low with a value of 0.745.

Table 7-2: RMSEC and RMSECV for the NPLS regression model for protein and starch content using the fused NIR and fluorescence data.

Variable	**Protein Content**	**Starch Content**
RMSEC	0.939	2.454
RMSECV	1.097	2.707
E (cal)	0.745	2.013

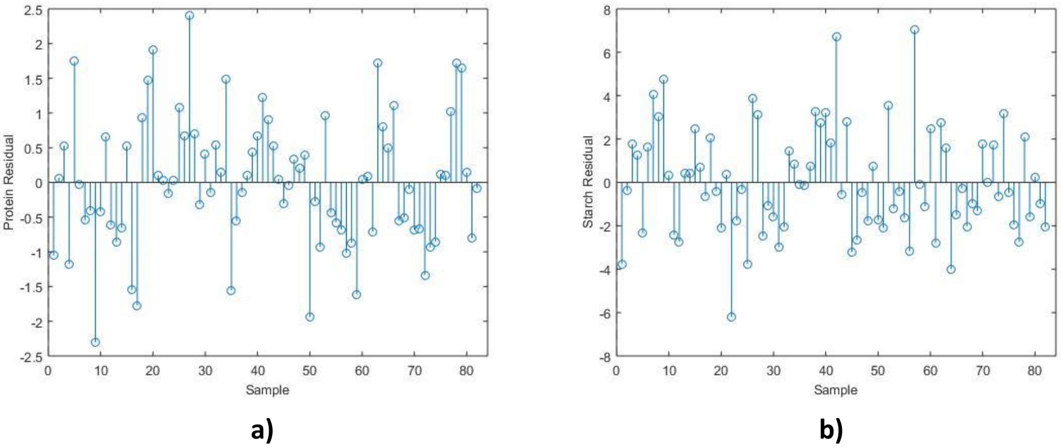


Figure 7-9: Absolute residuals for protein and starch content for the determination of outliers for the NIR and fluorescence fused data [a: protein content residuals, b: starch content residuals].

7.2.4 Data Fusion Model Comparison

The PLS regression model of the fused NIR and fluorescence data is used to predict the protein and starch content of all the samples. The same 82 samples are used in the calibration set and 20 samples used to validate the model as done for the NIR and fluorescence models. Figures 7-10 and 7-11 compare the model's predicted protein and starch content against the measured protein and starch content, respectively for both the calibration and validation sets.

The models were evaluated using the R^2 values. The calibration model for protein content is almost perfect having an R^2 value over 0.98. In Figures 7-10 and 7-11, it is apparent that the model for protein content is stronger than the model for starch content. This is due to the higher R^2 values for both calibration and prediction.

The goal of fusing the NIR and fluorescence spectra was to see if the fused data improved the predictability and robustness of the model. Table 7-3 compares the results of the fused data with the individual models for each spectroscopic technique.

Table 7-3: Comparison of NIR, fluorescence, and data fused models for protein and starch content of navy bean flour samples.

	Protein Content			**Starch Content**		
	NIR	**Fluorescence**	**Fused Data**	**NIR**	**Fluorescence**	**Fused Data**
RMSEC	1.152	1.769	0.939	2.437	3.538	2.454
RMSECV	1.189	1.867	1.097	2.579	3.647	2.707
R^2 (cal)	0.972	0.936	0.982	0.940	0.874	0.939
R^2 (val)	0.965	0.946	0.972	0.913	0.885	0.910
PC	3	3	5	3	3	5

The fused data actually performs better when comparing the value for protein content. Both R^2 values are higher and the root mean squared values are lower which results in the better model. This can be explained as both the NIR and fluorescence spectra can have spectral areas to determine protein content. The NIR spectra contained the absorbance values from the amide bonds contained in protein while the fluorescence spectra has both the tryptophan and unknown fluorophore attributed to protein content. The worst model for protein content came from the fluorescence spectra individually but is still a very reliable model to predict protein content.

The difference between the fused data and NIR models for starch content differ minimally. This is expected because the fluorescence data did not add any information as to the starch content in the samples since the two fluorophores did not directly relate to starch content. As discussed earlier, the starch content in the fluorescence model has an incredibly strong relationship to the predicted protein content. Once again the fluorescence model performed the worst in predicting starch content but is still a very good model.

The NIR and fluorescence models by themselves are both extremely good leaving little room for improvement by fusing the two data sets. Overall, there is only a minimal improvement in the protein content model when fusing the NIR and fluorescence data for predicting sample protein and starch content. While this is the case for a navy bean flour calibration and validation sets, the fused model may have a better correlation using the navy bean flour as the calibration model to predict the protein and starch content of a different sources of flour (Anibal *et al.* 2011; Dearing *et al.* 2011; Godinho *et al.* 2014).

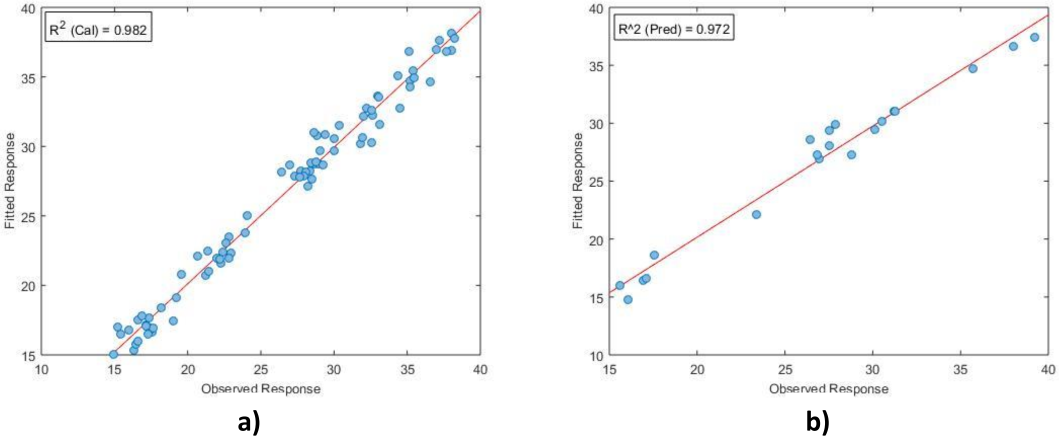


Figure 7-10: NPLS Regression model for protein content using NIR and fluorescence fused data [a: calibration model, b: validation model].

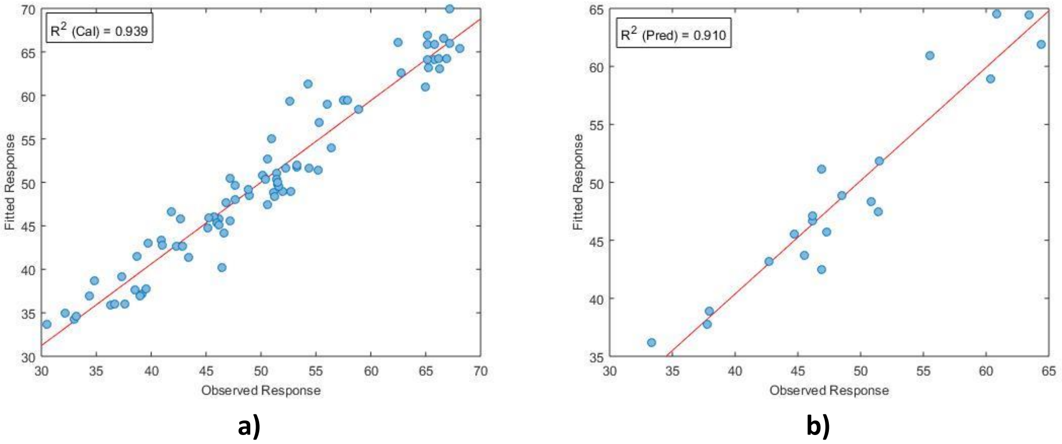


Figure 7-11: NPLS Regression model for starch content using NIR and fluorescence fused data [a: calibration model, b: validation model].

Chapter 8

Conclusions

The objective of this research was to establish the effects of different milling techniques on the solvent-free electrostatic separation process for navy bean flour as well as to develop a model based on near infrared and fluorescence data to determine protein and starch content of the protein- and starch-enriched fractions using multivariate methods (i.e. partial least squares regression). Acquisition of reproducible infrared and fluorescence data from powder samples was also an objective. The following list represents the conclusions from this work:

- Successful development of a method to obtain reproducible and reliable NIR and fluorescence data from compressed powder pellets.
- The pin milled navy bean flour had an average particle size almost three times smaller than the regular milled navy bean flour. This was likely the main factor contributing to the ability to produce a higher protein content in the protein-enriched fraction for the pin milled flour (40.7%) compared to the regular milled flour (32.5%) under optimal conditions. However, a much higher protein extraction under optimum conditions could be achieved for the regular milled flour.
- For regular milled navy bean flour the two fractions (protein- and starch-enriched) had a higher percentage of smaller particles than the raw flour. This might be attributed to disaggregation during the triboelectric charging process.
- Multivariate methods and pre-treatment techniques were compared for the NIR spectra of the navy bean flour fractions from separation to measure the protein and starch content.

The best method used MSC pre-treatment with PLS regressions and had R^2 values of prediction of 0.965 and 0.912 for protein and starch content, respectively.

- The models for protein and starch content using NPLS regression with Rayleigh scatter replaced by missing values was not as highly correlated as the NIR model. It was still a good model seeing as the R^2 values of prediction for starch and protein content of 0.946 and 0.885, respectively. Two fluorophores were observed in navy bean flour: one attributed to tryptophan and an unknown peak.
- The protein content model was better calibrated using the training set as well as providing a better prediction using the validation set for both NIR and fluorescence spectra. This was observed from the protein content model having smaller residual values and better correlations. While the NIR spectra have areas that can be attributed to bonds unique to starch, there are no fluorescent compounds. It was found that the starch model using the fluorescence EEM was highly correlated to the model's predicted protein content (R^2 of 0.978).
- Data fusion was achieved by combining the NIR and unfolded fluorescence spectra for the navy bean flour fractions. The individual approaches had undergone pre-treatment separately which involved SNV for the NIR spectra and reduction of data points for the unfolded fluorescence spectra. The best model for determining protein content used the fused data illustrating the value in this approach.

Chapter 9

Recommendations

- Further investigation into the unknown fluorophore. It has a shape and excitation/ emission pair that is similar to humic acid, which is ruled out because it is not found in navy bean flour, but further characterization of the flour could provide insight into the unknown fluorophore.
- Increase the number of measured components in the model particularly fat content and moisture content.
- Scanning electron microscopy of the raw and enriched fractions would help to confirm the disaggregation hypothesis developed to explain the size distribution phenomena observed for regular milled flour during triboelectric separation.
- Better design for keeping the distance between the optic probe and compressed flour pellet consistent for each sample during fluorescence data collection. This could have contributed to the fluorescence model not performing as well as the NIR model.
- Further work on the effects of milling on the electrostatic separation. Sieving the flour into fractions of differing size could provide insight into the separation of particular particle sizes.
- Further research using the electrostatic separation approach for different types of agricultural flour (e.g. soy, wheat, etc.) should be explored. The models for estimating protein and starch content of navy bean flour could be used as the calibration set for other flours providing insight into the robustness of the model.

References

Acharid, A., Rizkallah, J., Ait-Ameur, L., Neugnot, B., Seidel, K., Särkkä-Tirkkonen, M., Kahl, J. & Birlouez-Aragon, I., 2012. Potential of front face fluorescence as a monitoring tool of neoformed compounds in industrially processed carrot baby food. Food Science and Technology, 49(2), pp.305–311.

Andersen, C. & Bro, R., 2003. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. Journal of Chemometrics, 17(4), pp.200-215.

Anibal, C., Di Callao, M. & Ruisánchez, I., 2011. H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices. Talanta, 84(3), pp.829–833.

Asekova, S., Han, S., Choi, H., Park S., Shin, D., Kwon, C., Shanno, J. & Lee, J., 2016. Determination of forage quality by near-infrared reflectance spectroscopy in soybean. Turkish Journal of Agriculture and Forestry, 7, pp.45–52.

Balabin, R., Lomakina, E & Safieva, R.Z., 2015. Artificial neural network (ANN) approach to biodiesel analysis: analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. Fuel, 90(5), pp.2007–2015.

Bi, Y., Yuan, K., Xiao, W., Wu, J., Shi, C., Xia, J., Chu, G., Zhang, G. & Zhou, G., 2016. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. Analytica Chimica Acta, 909, pp.30-40.

Botelho, B., Oliveira, L. & Franca, A., 2017. Fluorescence spectroscopy as tool for the geographical discrimination of coffees produced in different regions of Minas Gerais State in Brazil. Food Control, 77, pp.25-31.

Braaksma A. & Schaap, D., 1996. Protein analysis of the common mushroom *Agancus bisporus*. Post Harvest Biology and Technology, 7, pp.119-127.

Bro, R., 1998. Multi-way analysis in the food industry. PhD Thesis, Amsterdam.

Bro, R. Nielson, H., Savorani, F., Kjeldahl, K., Christensen, I., Brunner, N. & Cawaetz, A., 2013. Data fusion in metabolomic cancer diagnostics. Metabolomics, 9, pp.3-8.

Bro, R., 1996. Multiway calibration. Multilinear PLS. Journal of Chemometrics, 10, pp.47-61.

Bro, R. & Kiers, H., 2003. A new efficient method for determining the number of components in PARAFAC models. Journal of Chemometrics, 17, pp.274-286.

Bro, R. & Vidal, M., 2011. EEMizer: Automated modeling of fluorescence EEM data. Chemometrics and Intelligent Laboratory Systems, 106(1), pp.86-92.

Callejón, R., Amigo, J. Pairo, E., Garmon, S., Ocana, J. & Morales, M., 2012. Talanta Classification of Sherry vinegars by combining multidimensional fluorescence, parafac and different classification approaches. Talanta, 88, pp.456-462.

Calvet, A., Li, B. & Ryder, A., 2014. A rapid fluorescence based method for the quantitative analysis of cell culture media photo-degradation. Analytica Chimica Acta, 807, pp.111-119.

Calvet, A., Li, B. & Ryder, A., 2012. Rapid quantification of tryptophan and tyrosine in chemically defined cell culture media using fluorescence spectroscopy. Journal of Pharmaceutical and Biomedical Analysis, 71, pp.89-98.

Chalus, P., Roggo, Y., Walter, S. & Ulmschneider, M., 2005. Near-infrared determination of active substance content in intact low-dosage tablets. Talanta, 66, pp.1294–1302.

Christensen, J., Norgaard, L., Bro, R. & Engelsen, J., 2006. Multivariate autofluorescence of intact Food Systems. Chemical Reiviews, 106(6).

Costa, G., Fernandes, D., Gomes A., Almeida, V. & Veras, G., 2016. Using near infrared spectroscopy to classify soybean oil according to expiration date. Food Chemistry, 196, pp.539–543.

Dastoori, K., Makin, B. & Tan, G., 2005. Measurement of the electrostatic powder coating properties for corona and triboelectric coating guns. Journal of Electrostatics, 63, pp.545–550.

Dearing, T., Thompson, W., Rechsteiner, C. & Marquardt, B., 2011. Characterization of crude oil products using data fusion of process Raman, infrared, and nuclear magnetic resonance (NMR) Spectra. Applied Spectroscopy, 65(2), pp.181–186.

Dramic, T., Bro, R., Zekovic, I., Dramicanin, T. & Dramicanin M., 2015. Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey. Food Chemistry, 175, pp.284–291.

Ely, D., Thommes, M. & Carvajal, M., 2008. Colloids and Surfaces A : Physicochemical and Engineering Aspects Analysis of the effects of particle size and densification on NIR spectra. Colloids and Surfaces A: Physicochemical and Engineering Aspects, 331, pp.63–67.

Favilla, S., Durante, C., Vigni, M. & Cocchi, M., 2013. Assessing feature relevance in NPLS models by VIP. Chemometrics and Intelligent Laboratory Systems, 129, pp.76–86.

Fernández, C., Callao, M. & Larrechi, M.., 2013. UV-visible-DAD and ^{I}H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS. Talanta, 117, pp.75–80.

Geladi P. & MacDougall D., 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. Applied Spectroscopy, 39(3), pp.491–500.

Godinho, M., Blanco, M., Gambarra Neta, K., Liao, L., Sena, M., Tauler, R. & de Oliveria, A., 2014. Evaluation of transformer insulating oil quality using NIR, fluorescence, and NMR spectroscopic data fusion. Talanta, 129, pp.143–149.

Grohganz, H., Gildemyn, D., Skibsted, E., Flink, J. & Rantanen, J., 2010. Towards a robust water content determination of freeze-dried samples by near-infrared spectroscopy. Analytica Chimica Acta, 676, pp.34–40.

Grosshans, H. & Papalexacndris M., 2016. Evaluation of the parameters influencing electrostatic charging of powder in a pipe flow. Journal of Loss Prevention in the Process Industries, 43, pp.83-91.

Grosshans, H. & Papalexandris, M., 2017. A model for the non-uniform contact charging of particles. Powder Technology, 305, pp.518–527.

Hall, M., 2009. Determination of starch, including maltooligosaccharides, in animal feeds: comparison of methods and a method recommended for AOAC collaborative study. Journal of AOAC International, 92(1), pp.42-49.

Harnly, J., Harrington, P., Botros, L., Jablonski, J., Chang, C., Bergana, M., Wehling, P., Dowrey, G., Potts, A. & Moore, J., 2014. Characterization of near-infrared spectral variance in the authentication of skim and nonfat dry milk powder collection using ANOVA-PCA, pooled-ANOVA, and partial least-squares regression. Journal of Agricultural and Food Chemistry, 62(32), pp.8060–8067.

Hemery, Y., Holopainen, V., Lampi, A., Lehtinen, P., Nurmi, T., Vieno, P., Edelmann, M. & Rouau, X., 2011. Potential of dry fractionation of wheat bran for the development of food ingredients, part II : electrostatic separation of particles. Journal of Cereal Science, 53(1), pp.9–18.

Hyun, C., Park, J., Ho, J. & Chun, B., 2008. Triboelectric series and charging properties of plastics using the designed vertical-reciprocation charger. Journal of Electrostatics, 66(11–12), pp.578–583.

Iban, C. & Ferrero, C., 2003. Extraction and characterization of the hydrocolloid from *Prosopis flexuosa* DC seeds. Food Research International, 36, pp.455–460.

Inculet, I., Castle, G. & Brown, J., 2017. Electrostatic separation of plastics for recycling. Particulate Science and Technology, 16, pp.91-100.

Jafari, M., Rajabzadeh, A., Tabtabaei, S., Marsolaid, F. & Legge, R., 2016. Physicochemical characterization of a navy bean (*Phaseolus vulgaris*) protein fraction produced using a solvent-free method. Food Chemistry, 208, pp.35–41.

Jong, S., 1993. SIMPLS : an alternative approach to partial least squares regression. Chemometricsand Intelligent Laboratory Systems, 18, pp.251-263.

Kamruzzaman, M., Makino, Y., Oshita, S. & Liu, S., 2015. Assessment of visible near-infrared hyperspectral imaging as a tool for detection of horsemeat adulteration in minced beef. Food Bioprocess Technology, 8, pp.1054–1062.

Khaleghi, B., Khamis, A., Karray, F. & Razavi, S., 2013. Multisensor data fusion: a review of the state-of-the-art. Information Fusion, 14, pp.28–44.

Korang-yeboah, M., Aktar, S., Siddiqui, A., Rahman, Z. & Khan, M., 2016. Application of NIR chemometric methods for quantification of the crystalline fraction of warfarin sodium in drug product Application of NIR chemometric methods for quantification of the crystalline fraction of warfarin sodium in drug product. Drug Development and Industrial Pharmacy, 42, pp.584-594.

Labair, H., Touhami, S., Tilmatine, A., Hadjeri, S., Medles, K. & Dascalescu, L., 2017. Study of charged particles trajectories in free-fall electrostatic separators. Journal of Electrostatics (2017), pp.1–5.

Lai, K., Lim, S. & Yeap, K., 2016. Modeling electrostatic separation process using artificial neural network (ANN). Procedia Computer Science, 91, pp.372–381.

Lang, C., 1958. Simple Microdetermination of Kjeldahl Nitrogen in Biological Materials. Analytical Chemistry, 30(10), pp.1692–1694.

Lee, B., Seo, Y. & Hur, J., 2015. Investigation of adsorptive fractionation of humic acid on graphene oxide using fluorescence EEM- PARAFAC. Water Research, 73, pp.242–251.

Lee, H., Chistie, A., Xu, J. & Yoon, S., 2012. Data fusion-based assessment of raw materials in mammalian cell culture. Biotechnology and Bioengineering, 109(11), pp.2819–2828.

Marengo, E., Bobba, M., Robotti, E. &Lenti, M., 2004. Hydroxyl and acid number prediction in polyester resins by near infrared spectroscopy and artificial neural networks. Analytica Chemica Acta, 511, pp.313–322.

Marquetti, I., Link, J., Guimaras Lemes, A., dos Santo Scholz, M., Valderrama, P. & Bona, E., 2016. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. Computers and Electronics in Agriculture, 121, pp.313–319.

Martelo-Vidal, M. & Vázquez, M., 2014. Determination of polyphenolic compounds of red wines by UV–VIS–NIR spectroscopy and chemometrics tools. Food Chemistry, 158, pp.28–34.

Matsusaka, S., Maruyama, H., Maruyama, T. & Ghadiri, M., 2010. Triboelectric charging of powders : a review. Chemical Engineering Science, 65(22), pp.5781–5807.

Megazyme: Total Starch Assay Procedure. (2016). Megazyme International Ltd.

Mizutani, M., Yasuda, M. & Matsusaka, S., 2013. Characterization of particle electrostatic charging in vibration and electric field. Chemical Engineering Transactions, 32, pp.2083–2088.

Mouazen, A., Kuang, B., De Baerdemaeker, J. & Ramon, H., 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma, 158, pp.23–31.

Murphy, K., Stedmon, C., Graeber, D. & Bro, R., 2013. Fluorescence spectroscopy and multi-way technique. PARAFAC. Analytical Methods. 5(23), pp.6557-6566.

Nadjem, A., Kachi, M., Bekkara, F., Zeghloul, T. & Dascalescu, L., 2017. Triboelectrification of granular insulating materials as affected by dielectric barrier discharge (DBD) treatment. Journal of Electrostatics, 86, pp.18–23.

Ndiokwere, C., 1984. Analysis of various Nigerian foodstuffs for crude protein and mineral contents by neutron activation. Food Chemistry, 14, pp.93–102.

Peiris, R.H. et al., 2013. Fouling control and optimization of a drinking water membrane filtration process with real-time model parameter adaptation using fluorescence and permeate flux measurements. Journal of Process Control, 23(1), pp.70–77.

Peiris, R., Budman, H., Moresoli, C. & Legge, R. 2012. Characterizing natural colloidal / particulate-protein interactions using fluorescence-based techniques and principal component analysis. Talanta, 99, pp.457–463.

Phong, D. & Hur, J., 2015. Insight into photocatalytic degradation of dissolved organic matter in UVA / TiO 2 systems revealed by fluorescence EEM-PARAFAC. Water Research, 87, pp.119–126.

Rajkovich, S., 2017. Advances in pin mill technology. Solids Processing (2017), pp.60-63.

Ranzan, C., Strom, A., Ranza, L., Trierweiler, L., Hitzmann, B. & Trierweiler, J., 2014. Wheat flour characterization using NIR and spectral filter based on ant colony optimization. Chemometrics and Intelligent Laboratory Systems, 132, pp.133–140.

Rinnan A. & Andersen, C., 2005. Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation-emission data. Chemometrics and Intelligent Laboratory Systems, 76, pp.91–99.

Rinnan, A., van den Berg, F. & Engelsen, S., 2009. Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry, 28(10), pp.1201–1222.

Ryan, P., Li, B., Shanahan, M., Leister, K. & Ryder, A., 2010. Prediction of cell culture media performance using fluorescence spectroscopy. Analytical Chemistry, 82(4), pp.1311–1317.

Schein, L., 1999. Recent advances in our understanding of toner charging. Journal of Electrostatics, 46, pp.29–36.

Schutyser, M., Pelgrom, P., van der Goot, A. & Boom, R., 2015. Dry fractionation for sustainable production of functional legume protein concentrates. Trends in Food Science & Technology, 45(2), pp.327–335.

Sharma, V. & Kalonia, D., 2003. Steady-state tryptophan fluorescence spectroscopy study to probe tertiary structure of proteins in solid powders. Journal of Pharmaceutical Sciences, 92(4), pp.890–899.

Shin, Y., Park, S., Yoo, J., Jeon, C., Lee, S. & Baek, K., 2016. A new approach for remediation of As-contaminated soil: ball mill-based technique. Environ Sci Pollut Res, pp.3963–3970.

Solano, M., Ekwaro-Osire, S. & Tanik, M., 2012. High-level fusion for intelligence applications using recombinant cognition synthesis. Information Fusion, 13, pp.79–98.

Tabtabaei, S., Vitelli, M., Rajabzadeh, A. & Legge, R., 2017. Analysis of protein enrichment during single- and multi-stage tribo-electrostatic bioseparation processes for dry fractionation of legume flour. Separation and Purification Technology, 176, pp.48–58.

Tabtabaei, S., Jafari, M., Rajabzadeh, A. & Legge, R., 2016a. Development and optimization of a triboelectrification bioseparation process for dry fractionation of legume flours. Separation and Purification Technology, 163, pp.48–58.

Tabtabaei, S, Jafari, M., Rajabzadeh, A. & Legge, R., 2016b. Solvent-free production of protein-enriched fractions from navy bean flour using a triboelectrification-based approach. Journal of Food Engineering, 174, pp.21–28.

Tripathi, M., Hassan, E., Yeuh, F. Singh, J., Steele, P. & Ingram, L., 2009. Reflection-absorption-based near infrared spectroscopy for predicting water content in bio-oil. Sensors and Actuators B: Chemical, 136(1), pp.20–25.

Vanarese, A., Alcala, M., Jerez Rozo, J., Muzzio, F. & Romanach, R., 2010. Real-time monitoring of drug concentration in a continuous powder mixing process using NIR spectroscopy. Chemical Engineering Science, 65, pp.5728–5733.

Wang, J., Smits, E., Boom, R. & Schutyser, M., 2015. Arabinoxylans concentrates from wheat bran by electrostatic separation. Journal of Food Engineering, 155, pp.29–36.

Wang, J., Suo, G. de Wit, M. Boom, R. & Schutyser, M., 2016. Dietary fibre enrichment from defatted rice bran by dry fractionation. Journal of Food Engineering, 186, pp.50–57.

Wise, B, Gallagher, N, Bro, R, Shaver, J, Windig, W, Koch, R, 2006. PLS Toolbox Manual. EigenVector Research.

Yan, M., Fu, Q., Li, D. Gao, G. & Wang, D., 2013. Study of the pH influence on the optical properties of dissolved organic matter using fluorescence excitation–emission matrix and parallel factor analysis. Journal of Luminescence, 142, pp.103–109.

Zeghloul, T., Benhafssa, A., Richard, G., Medles, K. & Dascalescu, L., 2016. Effect of particle size on the tribo-aero-electrostatic separation of plastics. Journal of Electrostatics (2016), pp.3–7.

Zenkiewicz, M., Zuk, T. & Markiewicz, E., 2015. Triboelectric series and electrostatic separation of some biopolymers. Polymer Testing, 42, pp.192–198.

Appendix A

Calibration Curves

A.1 Kjeldahl Calibration Curve

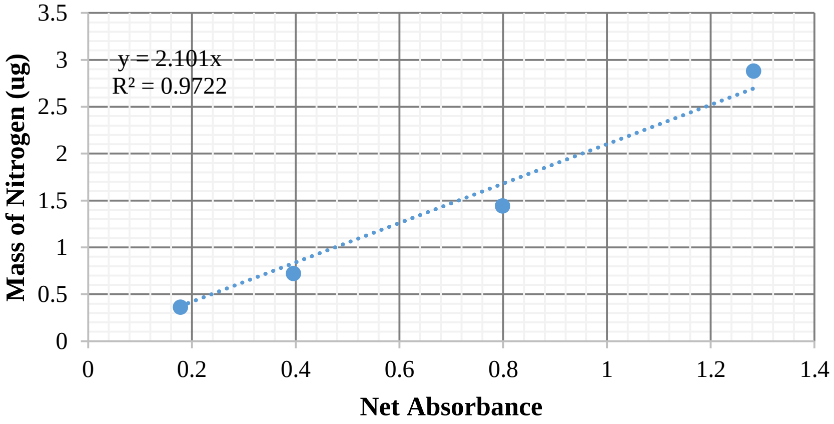


Figure A-1: Calibration curve for Kjeldahl digestion.

A.2 DNS Calibration Curve

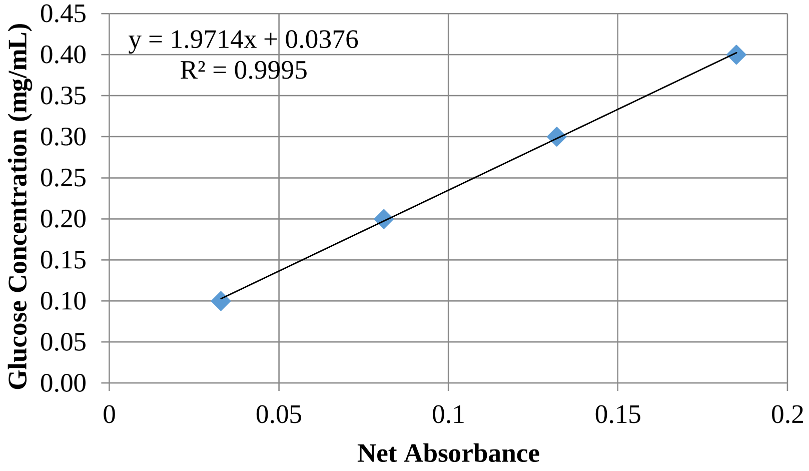


Figure A-2: Calibration curve for DNS acid assay.

Appendix B

Whey Protein Isolate Standard

B.1 NIR Standard

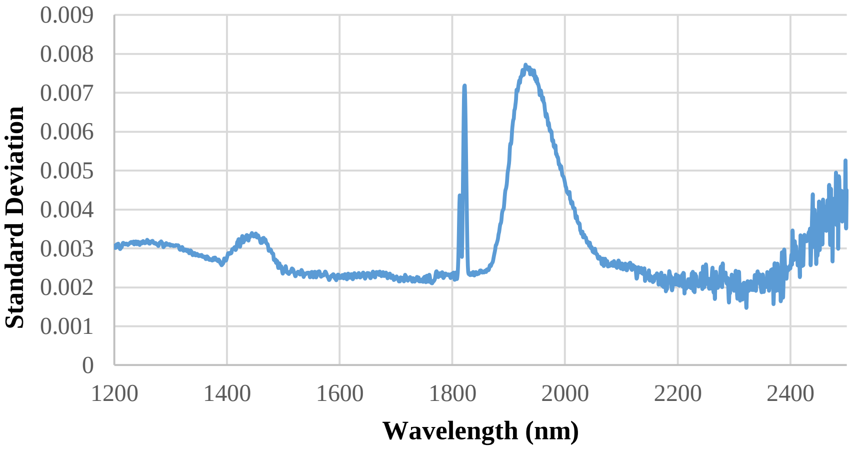


Figure B-1: Standard deviation of each NIR wavelength for 10 standard samples of whey protein isolate (~97%).

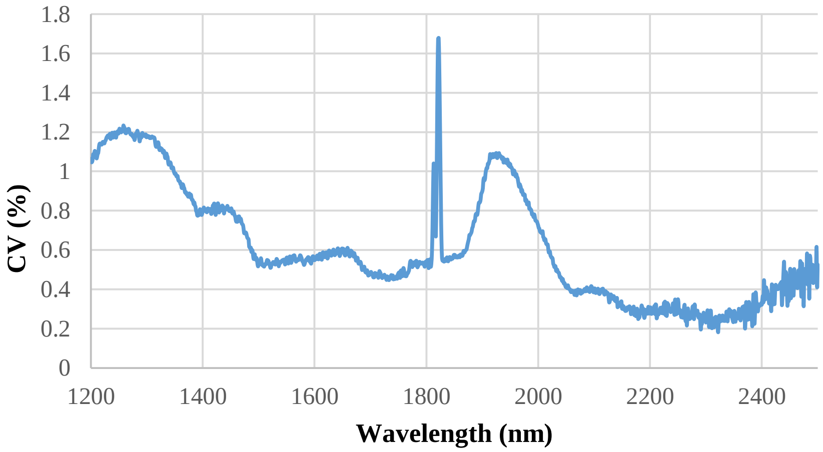


Figure B-2: CV of each wavelength for 10 standard samples of whey protein isolate (~97%).

B.2 Fluorescence Standard

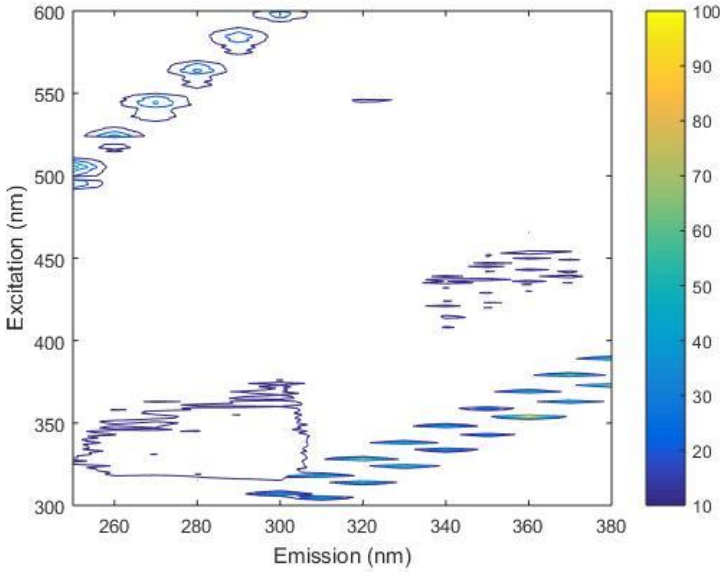


Figure B-3: Standard deviation of the fluorescence EEM for 10 standard samples of whey protein isolate (~97%).

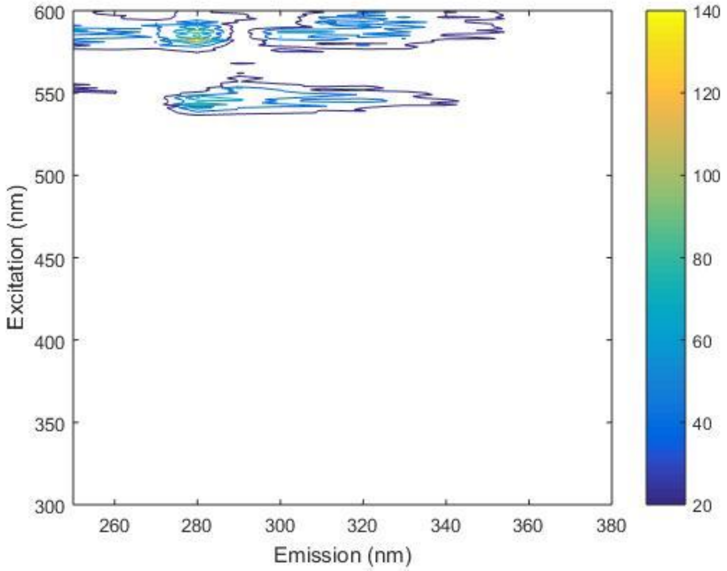


Figure B-4: CV of the fluorescence EEM for 10 standard samples of whey protein isolate (~97%).

Appendix C

Protein and Starch Content of Navy Bean Flour Samples

C.1 Calibration Set

Table C-1: Navy bean flour samples raw data in the calibration set.

Sample Number	Sample	Protein Content				Starch Content			
		Run 1	**Run 2**	**Average**	**Standard Deviation**	**Run 1**	**Run 2**	**Average**	**Standard Deviation**
1	Flour 1	28.60	27.80	28.20	0.57	55.83	49.62	52.73	4.39
2	July 20 Pb	30.50	26.90	28.70	2.55	47.94	44.49	46.22	2.44
3	July 20 Pm	33.40	31.00	32.20	1.70	36.63	38.07	37.35	1.02
4	July 20 Pt	38.00	38.10	38.05	0.07	32.61	33.32	32.97	0.50
5	July 20 S	27.00	26.90	26.95	0.07	53.93	48.45	51.19	3.87
6	July 20 P1	22.20	21.80	22.00	0.28	58.92	51.64	55.28	5.15
7	July 20 P23	23.70	18.80	21.25	3.46	49.26	52.66	50.96	2.40
8	July 20 P4	22.10	20.80	21.45	0.92	59.73	52.26	56.00	5.28
9	Aug 6 Pb	29.30	35.80	32.55	4.60	39.48	44.24	41.86	3.37
10	Aug 6 Pm	31.90	33.40	32.65	1.06	48.99	42.48	45.74	4.60
11	Aug 6 Pt	34.20	34.60	34.40	0.28	46.31	46.98	46.65	0.47
12	Aug 6 P1	23.40	22.50	22.95	0.64	54.91	53.80	54.36	0.78
13	Aug 6 P23	23.50	22.20	22.85	0.92	47.83	47.52	47.68	0.22
14	Aug 6 P4	22.40	22.10	22.25	0.21	50.61	49.81	50.21	0.57
15	Aug 10 Pb	29.47	30.60	30.04	0.80	39.63	42.15	40.89	1.78
16	Aug 10 Pm	32.02	34.20	33.11	1.54	42.17	48.25	45.21	4.30
17	Aug 10 Pt	33.89	35.10	34.50	0.86	45.94	45.95	45.95	0.01
18	Aug 10 P1	25.54	22.60	24.07	2.08	46.96	48.36	47.66	0.99
19	Aug 10 P4	22.58	18.73	20.66	2.72	50.41	52.48	51.45	1.46
20	Aug 16 Pb	29.53	28.14	28.84	0.98	41.26	45.61	43.44	3.08
21	Aug 16 Pm	32.39	31.69	32.04	0.49	41.91	42.67	42.29	0.54
22	Aug 16 Pt	35.95	34.92	35.44	0.73	48.53	44.36	46.45	2.95
23	Aug 16 P1	24.76	23.15	23.96	1.14	49.31	53.74	51.53	3.13
24	Aug 16 P23	23.81	20.95	22.38	2.02	45.18	45.10	45.14	0.06
25	Aug 16 P4	21.68	21.06	21.37	0.44	54.13	56.29	55.21	1.53
26	Nov 1 Pb	32.40	33.53	32.97	0.80	37.70	32.00	34.85	4.03
27	Nov 1 S	27.70	29.55	28.63	1.31	41.71	43.69	42.70	1.40
28	Nov 1 P1	23.70	21.94	22.82	1.24	58.47	54.35	56.41	2.91
29	Nov 1 P23	23.10	21.36	22.23	1.23	53.50	49.39	51.45	2.91

30	Nov 1 P4	24.10	21.19	22.65	2.06	54.82	51.78	53.30	2.15
31	Flour 4	30.17	26.56	28.37	2.55	51.10	52.81	51.96	1.21
32	Nov 8 Pb	34.24	31.81	33.03	1.72	37.82	40.59	39.21	1.96
33	Nov 8 Pmt	39.74	36.29	38.02	2.44	30.53	35.81	33.17	3.73
34	Nov 8 S	31.16	27.59	29.38	2.52	44.69	48.98	46.84	3.03
35	Nov 8 P1	20.71	17.27	18.99	2.43	68.19	65.01	66.60	2.25
36	Nov 8 P23	18.64	16.35	17.50	1.62	61.45	64.14	62.80	1.90
47	Nov 8 P4	19.98	18.52	19.25	1.03	65.26	65.01	65.14	0.18
48	Nov 11 Pb	31.74	32.34	32.04	0.42	40.57	38.92	39.75	1.17
39	Nov 11 Pmt	37.96	36.40	37.18	1.10	32.90	31.51	32.21	0.98
40	Nov 11 S	29.22	28.87	29.05	0.25	46.58	47.86	47.22	0.91
41	Nov 11 P1	19.04	20.16	19.60	0.79	57.43	57.67	57.55	0.17
42	Nov 11 P23	16.66	16.58	16.62	0.06	53.75	51.46	52.61	1.62
43	Flour 6	28.53	26.93	27.73	1.13	52.00	52.48	52.24	0.34
44	Nov 16 Pb	32.67	32.49	32.58	0.13	37.86	39.63	38.75	1.25
45	Nov 16 S	30.57	29.41	29.99	0.82	50.41	50.80	50.61	0.28
46	Nov 16 P1	17.09	17.23	17.16	0.10	68.09	68.11	68.10	0.01
47	Nov 16 P23	17.22	17.47	17.35	0.18	61.63	56.23	58.93	3.82
48	Nov 16 P4	18.95	17.44	18.20	1.07	65.81	65.89	65.85	0.06
49	Flour 7	28.87	28.00	28.44	0.62	51.72	48.47	50.10	2.30
50	Nov 18 Pmt	38.46	34.69	36.58	2.67	41.06	38.00	39.53	2.16
51	Nov 18 S	30.14	27.86	29.00	1.61	50.63	52.66	51.65	1.44
52	Nov 18 P1	17.66	17.43	17.55	0.16	62.87	62.17	62.52	0.49
53	Nov 18 P4	17.45	16.27	16.86	0.83	67.14	67.20	67.17	0.04
54	Nov 21 Pmt	35.98	34.40	35.19	1.12	35.09	37.50	36.30	1.70
55	Nov 21 S	28.61	29.95	29.28	0.95	45.59	48.77	47.18	2.25
56	Nov 21 P1	16.96	17.97	17.47	0.71	66.56	65.96	66.26	0.42
57	Nov 21 P23	16.19	16.45	16.32	0.18	56.29	52.33	54.31	2.80
58	Flour 9	28.30	28.67	28.49	0.26	51.10	49.77	50.44	0.94
59	Nov 23 Pb	31.53	32.11	31.82	0.41	46.42	45.91	46.17	0.36
60	Nov 23 Pt	37.15	36.77	36.96	0.27	33.67	35.11	34.39	1.02
61	Nov 23 S	28.82	27.30	28.06	1.07	49.97	52.46	51.22	1.76
62	Nov 23 P1	16.27	16.65	16.46	0.27	66.47	67.86	67.17	0.98
63	Nov 23 P23	16.27	14.23	15.25	1.44	59.32	56.41	57.87	2.06
64	Nov 23 P4	17.15	14.82	15.99	1.65	65.71	64.23	64.97	1.05
65	Flour 10	28.69	25.98	27.34	1.92	50.50	52.54	51.52	1.44
66	Nov 24 Pb	30.80	29.97	30.39	0.59	39.90	45.90	42.90	4.24
67	Nov 24 Pm	36.43	34.55	35.49	1.33	38.17	39.81	38.99	1.16
68	Nov 24 Pt	39.58	36.93	38.26	1.87	37.58	39.53	38.56	1.38
69	Nov 24 S	27.90	28.01	27.96	0.08	54.90	51.62	53.26	2.32
70	Nov 24 P1	16.88	16.39	16.64	0.35	65.63	64.61	65.12	0.72

71	Nov 24 P4	16.99	18.22	17.61	0.87	65.38	66.28	65.83	0.64
72	Nov 28 Pb	31.11	32.76	31.94	1.17	40.40	41.60	41.00	0.85
73	Nov 28 Pm	35.46	35.00	35.23	0.33	34.88	38.49	36.69	2.55
74	Nov 28 Pt	38.60	36.72	37.66	1.33	30.84	30.14	30.49	0.49
75	Nov 28 S	29.13	26.15	27.64	2.11	46.21	51.62	48.92	3.82
76	Nov 28 P23	15.58	14.36	14.97	0.86	65.91	66.36	66.14	0.32
77	Nov 28 P4	17.02	13.88	15.45	2.22	64.09	69.79	66.94	4.03
78	Flour 12	27.70	25.14	26.42	1.81	50.66	50.59	50.63	0.05
79	Dec 1 Pm	34.36	35.93	35.15	1.11	37.15	38.07	37.61	0.65
80	Dec 1 S	27.92	29.59	28.76	1.18	50.81	46.94	48.88	2.74
81	Dec 1 P1	16.48	18.09	17.29	1.14	64.99	65.29	65.14	0.21
82	Dec1 P4	16.71	17.63	17.17	0.65	64.54	65.96	65.25	1.00
				AVG	**1.14**			**AVG**	**1.75**

C.2 Validation Set

Table C-2: Navy bean flour samples raw data in the validation set.

Sample Number	**Sample**	**Protein Content**				**Starch Content**			
		Run 1	**Run 2**	**Average**	**Standard Deviation**	**Run 1**	**Run 2**	**Average**	**Standard Deviation**
83	Aug 6 Pds	29.70	30.50	30.10	0.57	47.59	44.77	46.18	1.99
84	Flour 2	24.72	28.20	26.46	2.46	53.04	48.68	50.86	3.08
85	Aug 10 Pds	27.83	27.90	27.87	0.05	46.70	47.02	46.86	0.23
86	Aug 10 P23	22.90	23.84	23.37	0.66	46.84	45.42	46.15	1.00
87	Aug 16 Pds	27.99	27.02	27.51	0.69	45.50	49.11	47.31	2.55
88	Flour 3	27.80	26.14	26.97	1.17	51.33	51.68	51.51	0.25
89	Flour 5	28.18	26.85	27.52	0.94	48.28	45.44	46.86	2.01
90	Nov 11 P4	18.07	17.00	17.54	0.76	62.52	59.20	60.86	2.35
91	Nov 16 Pmt	39.72	38.78	39.25	0.66	31.88	34.78	33.33	2.05
92	Nov 18 Pb	31.55	30.90	31.23	0.46	42.77	48.25	45.51	3.87
93	Nov 18 P23	17.59	16.31	16.95	0.91	62.73	58.02	60.38	3.33
94	Flour 8	27.97	25.70	26.84	1.61	47.18	49.87	48.53	1.90
95	Nov 21 Pb	30.87	31.65	31.26	0.55	44.38	41.09	42.74	2.33
96	Nov 21 P4	17.70	16.47	17.09	0.87	63.59	65.24	64.42	1.17
97	Nov 23 Pm	34.80	36.61	35.71	1.28	38.74	37.14	37.94	1.13
98	Nov 24 P23	16.66	14.51	15.59	1.52	64.34	62.45	63.40	1.34
99	Flour 11	28.18	29.36	28.77	0.83	49.18	53.63	51.41	3.15
100	Dec 1 Pb	30.34	30.73	30.54	0.28	43.34	46.15	44.75	1.99
101	Dec 1 Pt	37.25	38.72	37.99	1.04	37.86	37.68	37.77	0.13
102	Dec 1 P23	16.81	15.29	16.05	1.07	52.69	58.31	55.50	3.97
				AVG	0.92			AVG	1.99