# Using Near-Infrared Reflectance Spectroscopy (NIRS) for Qualitative determination of undesirable chemical component of high nitrogen content in protein raw material used for fish feed

Master Thesis in Chemometrics

Satvir Kaur Bhatti

Department of Chemistry, University of Bergen

June 1. 2018

# Acknowledgement

First, I would like to sincerely thank my supervising Prof. Bjørn Grung. His advice, knowledge, encouragement and support helped me tremendously to gain an understanding of the subject, and helped me to complete this thesis successfully.

I owe my deepest gratitude to Skretting ARC for providing me with the opportunity to carry out the work related to this thesis and project, using their well-equipped laboratory and facilities.

I fall short of words to thank Mrs Paivi Teivainen-Lædre Lab manager from Skretting ARC, to provide me with the opportunity to work on this thesis project, and for providing me with all the necessary tools and software required to complete this project and thesis.

I would specially like to thank Mr Ørjan Breivik my internal supervisor and Global NIR responsible / senior researcher from Skretting ARC, for believing in me. He took precious time from his busy schedule to make himself available whenever required. He attended meetings regarding the project and held guidance sessions for me. His guidance, professional expertise and advice on the subject greatly helped me to gain an in-depth understanding of the subject and helped me complete my thesis satisfactorily.

I am thankful to my husband Ranjodh Singh for all the support and encouragement in this period. You are always supportive and inspire me to do my best.

Finally, I would like to thank my kids Ranvir and Rajvir for sleeping on time and letting me write my thesis. You guys are wonderful!!

Thank you

Satvir Kaur Bhatti

# Abstract

Food safety and authenticity are important issue. Ingredients presenting high value are the most vulnerable for adulteration as the common practice is to replace original substance partially with cheap and easily available substance for economic gains. Authentication is also of concern to manufacturers who do not wish to be subjected to unfair competition.

Fishmeal has been the major source of protein in feeds for farmed fish. Due to increase in the growth of aquaculture production and limited availability of FM, alternative protein sources such as plant proteins (PP) are used. Wheat gluten is a PP source that has given promising results. Wheat gluten is made by washing wheat flour dough with water until all the starch granules and soluble fiber have been removed. It is a high protein raw material with good digestibility and interesting amino acid profile in addition to be used for its binding property. Due to these qualities use of wheat gluten as plant protein source has considerably increased in aquaculture feeds.

The aim of this study is to use NIRS and chemometric tools for the early discrimination of adulterated wheat gluten samples from pure wheat gluten samples. A SIMCA model was developed to discriminate between adulterated and unadulterated samples. SIMCA model showed 100 % classification at adulteration level of 3000 ppm .Thus, NIRS together with SIMCA model represent an attractive option for quality screening without sample pretreatments.

.

# List of Abbreviations and Notations

| | |
|---|---|
| CV | Cross validation |
| EMA | Economically Motivated Adulterants |
| EMSC | Extended multiplicative signal correction |
| LV | Latent variables |
| NIRS | Near infrared spectroscopy |
| NPN | Non- protein Nitrogen |
| PCA | Principal component analysis |
| PCA | Principal components |
| PLS | Partial least square |
| PP | Plant Proteins |
| RSD | Relative standard deviation |
| SIMCA | Soft independent modelling of class analogy |
| Wg | Wheat gluten |

# Contents

# 1 Introduction

## 1.1 Background

Food adulteration is the process of replacing original substance partially with cheap substance and thereby lowering or degrading the quality and effecting nutrients like protein, fat, carbohydrates, vitamins and others, that are important for normal growth. Protein is a high value ingredient since it plays a vital role in a number of important functions such as, catalyzing metabolic reactions, DNA replication and intracellular transport from one location to other. Protein is thus most vulnerable for adulteration. Proteins are large complex biomolecules consisting of one or more long chains of amino acids. Amino acids are organic compounds containing amino, carboxyl group and side chain (R-group) specific to each amino acid Figure 1.1

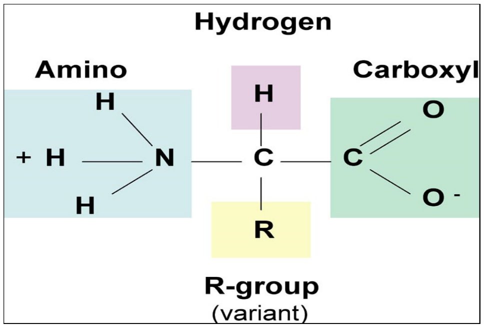


Figure 1.1: Amino Acid Structure

A non-protein nitrogen (NPN) compound is a term used for substances that have element nitrogen in them, but are not protein. For example melamine, cyanuric acid and urea amongst others. Due to low cost of NPN compounds, compared to plant and animal protein, they may be added to raw materials to artificially increase crude protein value. Such substances are called Economically Motivated Adulterants (EMA) and are intentionally added for financial gains. EMA present a challenge to the food industry and regulatory authorities because they are deliberate acts that are intended to evade detection. Journal of Food protection have reviewed some of EMA incidents [1]. In March 2007 contaminated pet food lead to the deaths of a number of cats in North America. This had prompted pet food recall in North America. In China in November 2008 there was a major food safety incident involving milk and infant formula adulteration causing the death of six infants from kidney damage and kidney stones. Due to these and other similar incidents, EMA has become a crucial safety issue for the food industry. EMA incidents reveal gaps in quality assurance testing methodologies that are exploited for financial gains.

As discussed in the study conducted by Phromkunthong [2] , inclusion of EMA in fish feed have shown harmful effects on fish and consumption of such fish may be hazardous to human health. An eight-week feeding trial demonstrated, that the fish fed on feed containing EMA grew less, utilized less feed and performed poorly. Fish fed on the adulterated feed also exhibited symptoms and defects like anorexia, sluggish swimming behavior, fin erosion, darkening of skin and high mortality. Food safety crises have aroused the need for a sensitive, reliable and rapid procedure for detection of possible adulterants. The standard protein determination assay, for example Kjeldahl method, measures total nitrogen in the samples and cannot differentiate between protein nitrogen and non-protein nitrogen. Hence, some producers for economic gains add NPN compounds and try to make their product seem to contain more protein than it actually does. European food and safety authority [3] in 2010 has set a maximum permitted concentration for NPN in food and feed at a level of 2.5 mg/kg, these are limits for low level of contamination. However, to make profit by EMA, much higher level of NPN needs to be added. Example ; for 2% addition of NPN to a raw material , the nitrogen content of the resulting mixture is increased by approximately 1.3% and the apparent protein content would be increased by over 8% assuming a nitrogen to protein ratio of 6.25 [4].

The traditional and novel detection methods like Gas Chromatography Mass-Spectrometry (GC-MS), High Performance Liquid Chromatography (HPLC), Capillary electrophoresis (CE), Nuclear magnetic resonance spectroscopy (NMR), Enzyme Linked Immunosorbent Assays (ELISA), Nanoparticle based sensors amongst others are very sensitive, but destructive, time consuming and require highly trained analysts. In addition they are costly as high-tech instruments are required [5] . New approaches based on biomimetic sensors, vibrational spectroscopy, Hyperspectral and Multispectral imaging (HIS-MSI) are being explored as rapid and non-destructive techniques for determination of authenticity and quality [6, 7].

Near Infrared Spectroscopy (NIRS) is a vibrational spectroscopy technique applied in areas such as nutrition and authenticity in aqua and agro culture. NIRS is becoming an important tool due to non-destructive capabilities, speed, reproducibility and ease of implementing this technology into an industrial set-up. NIRS in combination with chemometrics can be used to discriminate between fishmeal, soya meal and meat meal samples [8]. NIRS can be used to discriminate  different species of fishmeal batches [9]. NIR spectroscopy to detect adulteration in soybean meal using multivariate calibration models has been demonstrated by Haughey [10].

## 1.2 Objective

The main objective of this study is to devise a method using near-infrared spectroscopy (NIRS) and chemometrics to detect the presence of NPN compound in wheat gluten that is used as a source of plant protein in fish feed. Multivariate statistical tools provide pattern recognition techniques that allow adequate differentiation to be made between authentic and unauthentic wheat gluten samples. Additional objective is to investigate the extent of adulteration which can be identified using PLSR model.

# 2 Theory

## 2.1 Spectroscopy

### 2.1.1 Near Infrared Spectroscopy

Near-infrared spectroscopy (NIRS) is a vibrational spectroscopy method that measures absorption in the near-infrared region of the electromagnetic spectrum, defined as wavelengths from approximately 700 to 2500 nm. The basic principal of NIRS is based on vibrational energy, which results in periodic displacement of atoms from their equilibrium state. When a sample is irradiated, molecules in the sample absorb light and they vibrate accordingly to their selective vibrating frequencies giving rise to a spectrum. The NIR region is characterized by overtone and combination bands of fundamental vibrations of –CH, -NH, -OH and –SH functional groups. The information in the NIR spectra is repeated through successive overtones and combinations. The intensity of bands involved become weaker towards shorter wavelengths. The weaker intensities in the NIR region mean that solid samples need no dilution and non-linearity effects due to strong absorption are less likely [11]. Interaction of near infrared radiation with solid particles give rise to refraction, transmittance, absorption, and scattering effects as shown in figure 2.1 [11].

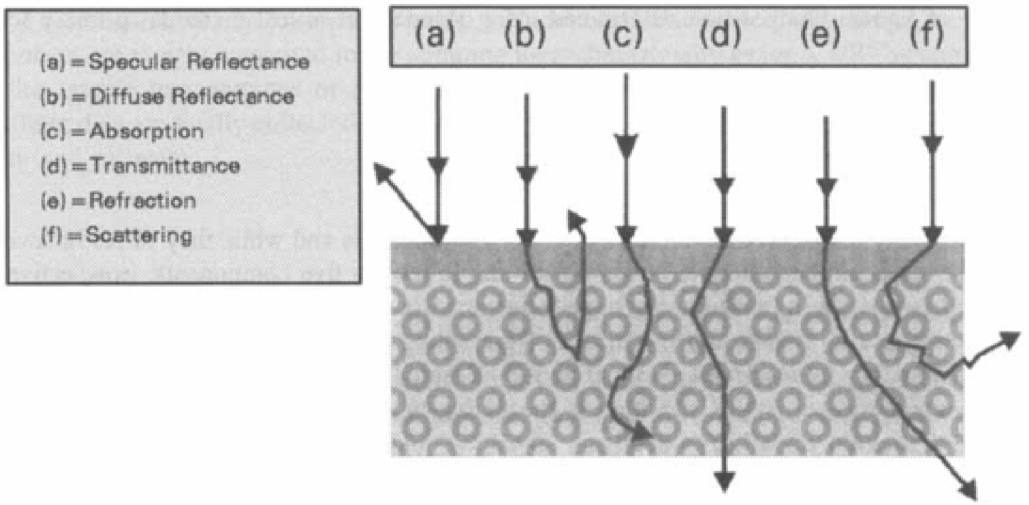


Figure 2.1. Interaction of near infrared radiation with solid particles in a sample

According to Beer-Lambert law, the amount of light absorbed (A) by the sample is directly proportional to the concentration (c) of the analyte, molar absorptivity (a) and path length (b).

The path length of radiation is maintained constant in absorption spectroscopy, but is affected by scattering of light for transmittance and reflectance spectroscopy. Scattering occurs when radiation transmitted through the surface is diffused by refraction, reflection and diffractions. The concentration and absorbed energy relation for NIRS region further involves overlapping of spectral bands from different constituents present in the sample, hence NIRS is an empirical technique and needs to be calibrated using standard chemical methods. NIRS is a simple, rapid, nondestructive technique that provides several parameters from one analysis, and hence cost effective compared to wet chemistry methods. It is a nondestructive technique and requires no sample preparation with hazardous chemicals, solvents or reagents. The instrument is safe and easy to use [11].

### 2.1.2 NIR Instrumentation

The basic NIR instrument configuration is either transmittance or reflectance figure 2.2. Irrespective of the configuration, both types consist of the following five components; source of energy, wavelength discrimination, sample holder or cup, detector, and signal processor. The common source of energy is tungsten filament lamp since it emits light from 320 to 2500 nm. Filters or monochromators are used for wavelength discrimination. Filters (usually between six and nineteen) are mounted on a rotating flat disc allowing radiation from the lamp to pass sequentially through each filter whereas monochromators scan the whole wavelength range by using a prism or grating as dispersing medium. Detection of NIR radiation occurs photo electrically. The incident photons change the electron state of the photosensitive material of detector, thereby producing an electrical impulse as detector output. To minimize scattering effect detectors are placed near the sample at 45 degree. The signals from the detector are amplified and readout as spectrum [11].

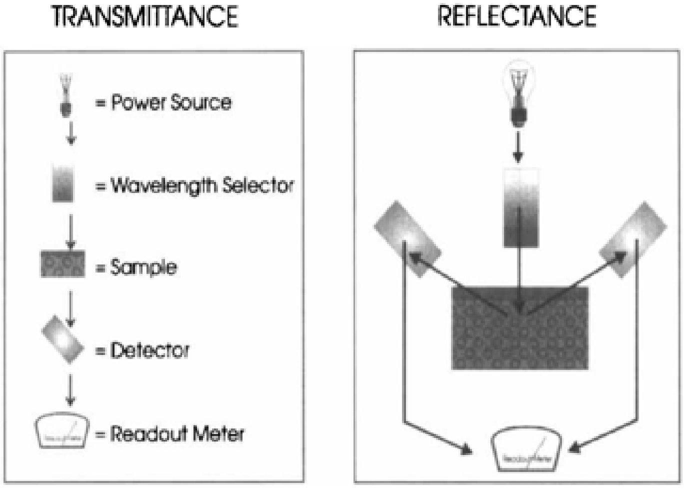


Figure 2.2: NIR instrument configuration for transmittance and reflectance

## 2.2 Multivariate Data Analysis

NIR spectra are complex and possess broad overlapping and combination of NIR absorption bands that require special mathematical techniques for data analysis. Modern near-infrared technology relies heavily on computers for its ability to acquire data from instruments and facilitate data analysis. Multivariate data analysis techniques can been applied effectively for explorative analysis, discrimination and classification or regression and prediction. The choice of technique depends on the goals of the analysis [12].

### 2.2.1 Spectral Preprocessing Techniques

Experimental and instrumental effects that are not related to sample compositional differences make sample comparison difficult. When samples are analyzed by either reflectance or transmittance NIRS, uncontrolled variations in light scattering is a dominating artifact. The spectra obtained contains noise and background information in addition to sample information. The undesired scattering variation is due to physical variation in the sample, such as particle size, sample surface, sample packing etc. The goal of data pre-treatment is to minimize variability unrelated to the property of interest. When analyzing spectral data, it is common to try out different pre-treatment and their combinations. Pre-treatment methods should be used with care as they can reduce signal of interest [13].

Derivatives are commonly used to remove unimportant signal from samples [14]. Derivatives are a form of high pass filter and are often used when high frequency features contain signal of interest. This method should only be used when the variables are strongly related to each other and the adjacent variables contain similar or correlated signal. The simplest form is first derivative, in which each variable is subtracted from its neighboring variable, to remove the signal that is similar and leaves the part of signal that is different. The first derivative thus removes any offset from the sample and deemphasizes low frequency signals. A second derivative is calculated by repeating the process and there by further accentuates higher frequency features. Since differentiation emphasizes higher frequencies, it also tends to accentuate noise and hence some form of smoothing is required along with differentiation.

Smoothing improves the signal to noise ratio by attenuating high frequency signals. Undersmoothing will not remove any noise, whereas oversmoothing will reduce the signal intensity and resolution. The optimum smoothing function depends on peak widths and noise characteristics [15]. Most common methods for smoothing are moving average or Savitzky-Golay smoothing [16]. In moving average (MA), a fixed number of data points are selected, their ordinates are added and then divided by the number of data points selected to obtain the average value. The number of data points selected is called window. The spectral data is smoothed by moving the window along the spectrum and by successively replacing each data point with a new point through entire dataset. Running median smoothing (RMS) operates in similar way but calculates median rather than mean over a window. Better noise reduction may be obtained by selecting more number of points, but this can lead into distorted signals as show in the figure 2.3 [15].

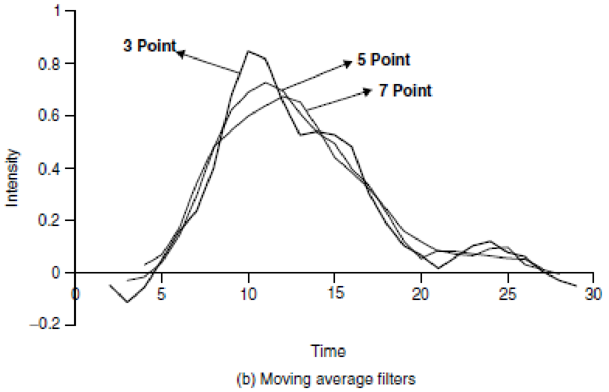


Figure 2.3: Change in signal with increase in window size

Savitzky - Golay is the most commonly used polynomial smoothing method to improve signal to noise ratio. Least square method is used to fit set of data points through a polynomial function to a data in a moving window. One needs to decide the number of points used to calculate the polynomial (window size) and the degree of polynomial fitting [15].

Multiplicative Scatter Correction (MSC) [13, 17] in its basic form was first introduced by Martens et al. (1983). It is a preprocessing technique used to remove non-linearity in the data caused by scattering from particle size of the sample. MSC deal with light scattering to compensate for multiplicative (amplification) and additive (offset) effect in the raw spectral signal [13]. It assumes that the light scattering for each sample is estimated relative to that of reference spectrum. Each spectrum is then corrected so that all samples appear to have the same scatter level as the reference spectrum. The reference spectra can either be a pre-defined reference or the average spectra over a set of samples (e.g. the calibration set). The average over a set of samples is normally used as it is difficult to obtain one appropriate reference spectrum [13]. MSC model for each individual spectrum comprises of two steps [13]

1. Estimation of the corrected coefficient

$$x_{org} = b_0 + b_{ref,1}\, x_{ref} + e \qquad (2.1)$$

2. Corrected spectra

$$x_{corr} = (x_{org} - b_0)/\ b_{ref,1} \qquad (2.2)$$

Where $X_{org}$ is one original sample spectra measured by NIR instrument, $X_{ref}$ is the reference spectrum, e corresponds effects that cannot be modelled in $X_{org}$ , $X_{corr}$ is the corrected spectra and $b_0$ and $b_{ref,1}$ are scalar parameters estimated by least square and differ for each sample.

Extended multiplicative signal correction (EMSC) is a modification of the MSC method to include wavelength corrections [13, 18]. With EMSC it is possible to estimate and separate multiplicative physical effects ( path length, sample thickness ,light scattering, etc. ) from additive physical effects (baseline variation ,temperature shifts , etc. ) and additive chemical effects (absorbance of analytes and interferants) [19]

Standard Normal Variate (SNV) [13, 20] is a preprocessing technique used for scattering correction. The signal correction concept behind SNV is same as for MSC except that common reference signal is not required. SNV transformation centers each spectrum and then scales it by its own standard deviation

$$X_{corr} = (X_{org} - a_0)/ a_1 \qquad (2.3)$$

Where $X_{corr}$ is the corrected spectra, $X_{org}$ is the original sample spectra, $a_0$ is the average value of the sample spectrum to be corrected and $a_1$ is the standard deviation of the sample spectrum.

### 2.2.2 Principal Component Analysis (PCA)

PCA is one of the most important multivariate explorative data analysis tool. PCA is a bilinear modeling techniques that provides a visual approach to identify patterns in data, outlier detection, variable selection, classification and dimension reduction. The possibility of using PCA for classification forms the basis for the classification method called SIMCA (Soft Independent Modelling of Class)[21]. PCA is also called as projection method as it uses information from original variables and projects them onto a smaller number of latent variables called Principal Components (PC). Each PC explains certain amount of information present in the original data. First PC stretch out in the direction of most variance, the next PC is orthogonal to this axis and has the direction where there is second most spread of variance. Thus, the first PC explains greatest amount of information in the data set and each subsequent PC explains less or remaining information than the previous one. The matrix X of the NIR spectral data has sample as rows and wavelength as columns and can be decomposed by PCA into a product of scores (T) and loading ($P^T$) matrix as illustrated in equation 2.4. For the loading matrix ($P^T$) superscript T implies transposition of column into row vector. E is residual matrix .Thus , E is the part of X that is not explained by the product $TP^T$ [12]. E is a good measure of “lack-of-fit” that describes how close the model is to the original data.

$$X= TP^T + E = \text{Structure} + \text{Noise} \qquad (2.4)$$

The score and loading plots are normally constructed using PC1verses PC2, as they explain the largest variance in the data set. The score plot of PC1 verses PC2 is shown in figure 2.4 [12]. The score plots reveals patterns or groupings of objects. On a score plot, objects that are closely clustered behave similar whereas objects that are diametrically opposite are negatively correlated [15]. The loading plots describe variable correlations. On a loading plot, variables that are close have high correlation whereas variables on opposite side of origin have negative correlation.

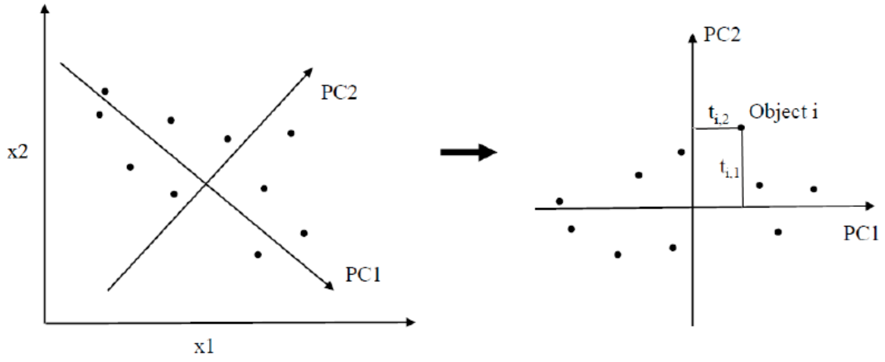


Figure 2.4: The plot to the left is of first and second principal components. The plot to right is a score plot illustrating the coordinates of the object i on PC1 and PC2.

NIR spectra consists of data collected at hundreds to thousands of different variables (wavelengths). Moreover, for NIR these variables (wavelengths) are highly correlated, meaning that the measured absorbance at two or more wavelengths are not independent of each other. This is referred to collinearity or multicollinearity and can pose problems with linear regression models [12]. To handle multicollinearity and to get a better overview of the data , it is necessary to reduce the amount of data [20]. Principal component analysis (PCA) among others is the most commonly used method for dimension reduction of the NIR spectra matrix [20]

### 2.2.3 SIMCA Classification

Soft independent modeling of class analogy (SIMCA) [22] , is a supervised pattern recognition class modeling technique. In class modeling the focus is on modeling the similarity between the samples belonging to a particular class. SIMCA classification algorithm is based on Principal Component Analysis (PCA). PCA is performed on each class to create a separate bilinear model for each group in the training set. The number of PCs needed to describe each group is usually different; too many PCs will add noise whereas few components can distort the information contained in the data. Hence, it is important to optimize the number of PCs retained for each class PC model. Cross-validation is used to find the number of PCs necessary to describe the data[22]. The variance that explains data is called model variance of the class model whereas the residual variance describes noise in the data and is part of the PCs not included in the model. Since SIMCA is based on PC model**s**, it is sensitive to the quality of the data used to create PC models. Parameters such as modeling power and discriminatory power are used to assess the quality of the data. Modeling power is a measure of how well a variable helps the PCs to model variation and has value between 0 and 1. Modelling power [15] close to 1 means that the variable is mostly accounted for by the model. A value close to zero indicates that the variable has a variation pattern distinct from the PCs, and such a variable should **be** deleted.

Modelling power of each variable for each separate class is given by equation 2.5

$M_{j\,=}\,1- S_{jresid}/ S_{jraw}$ (2.5)

Where $S_{jraw}$ is the standard deviation of the variable in the raw data and $S_{jresid}$ the standard deviation of the variable in the residuals. Discriminatory power [15]describes how well a variable helps $PC_s$ to discriminate between two groups; it is a positive number equal to or greater than one. A value close to one indicate that the variable has no ability to distinguish where as a value greater than three indicates good separation test is used to compare the residual variance of unknown sample with mean residual variance of the class model.

The equations below mathematically describe detailed procedure of SIMCA [23]

$$s_0^K = \left[ \sum_{j=1}^{m} \sum_{i=1}^{n_K} \frac{\left(e_{ij}^K\right)^2}{(n_K - p_K - 1)(m - p_K)} \right]^{1/2} \qquad (2.6)$$

Where $s_0^K$ is the mean residual standard deviation of the training set for class K, $n_k$ is the number of objects, $p_k$ is the number of significant $PC_s$ in class K, m is the number of variables and $e_{ij}$ is the residual.

$$s_u^K = \left[ \sum_{j=1}^{m} \frac{\left(e_{uj}^K\right)^2}{(m - p_K)} \right]^{1/2} \tag{2.7}$$

The residual standard deviation of the unknown spectrum, $s_u^k$ is given by equation 2.7 and is calculated using $e_{uj}$ value.

Comparison of the relative standard deviation (RSD) for the unknown ($s_u^k$ ) with the mean RSD for the model $s_0^K$ gives a direct measure of its similarity to the subset model. F test statistics is used for the comparison of $s_u^k$ and $s_0^K$

If the F-value is larger than the critical F-value at a given level of significance, it can be concluded that the distance from class K is significantly larger, i.e. the sample does not belong to the class K.

$$F = \frac{\left(s_u^K\right)^2}{\left(s_0^K\right)^2}$$

$$\left(s_{\text{lim}}^K\right) = \left(s_0^K\right) F^{0.05} \tag{2.8}$$

### 2.2.4 Univariate and multivariate calibration models

Calibration modeling involves using empirical data and prior knowledge for predicting concentration of the unknown samples. Univariate calibration model or simple linear regression model consists of dependent variable (y), independent variable (x), coefficient term ($b_0$ and $b_1$) and unexplained variance in the dependent variable which is given by error (e), as shown in the equation 2.9. The coefficient terms $b_0$ and $b_1$ are found using least square principle, as given in the equation 2.5 [12].

$$y= b_0+ b_1x+ e \qquad (2.9)$$

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{SS(xy)}{SS(x)}$$

$$b_0 = \frac{\sum y - (b_1 \cdot \sum x)}{n} = \bar{y} - (b_1 \cdot \bar{x}) \qquad (2.10)$$

NIRS is non-selective, meaning there is no single wavelength that provides sufficient information and the absorbance at all wavelengths are effected by physical and chemical property of the sample. The selectivity problem in NIRS can be solved by using several wavelengths i.e. having number of independent variables. However, a simple linear regression model cannot be applied when there is more than one independent variable. Multivariate calibration or multiple linear regression model (MLR) is used to solve this situation. A multivariate calibration model is illustrated in equation 2.11.

$$y= b_0+b_1x_1+b_2x_2+\ldots\ldots b_kx_k+ e \qquad (2.11)$$

The above equation can be written in the matrix form as equation 2.12

$$y= X\,b + e \qquad (2.12)$$

The vector of regression coefficients b ($b_0$, $b_1$,$b_2$…$b_k$ ) is found by least squares fitting so as to minimize the sum of squares residuals as given in equation 2.13

$$b= (X^TX)^{-1}X^Ty \qquad (2.13)$$

In case of strong collinearities in X variables, ($X^TX$) is no longer a non-singular or full rank matrix and inverse is not possible [24]. This a drawback of MLR.

Partial least square regression (PLSR) is a multivariate calibration technique used to predict dependent variables from independent variables. PLSR is a dimension reduction technique that uses original variables to calculate number of latent variables called factors. PLSR can thus be used to handle collinearity issues with X variables. It uses covariance between independent variables in data matrix X and dependent variables in response matrix Y. Thus both data matrix X and response matrix Y are decomposed into product of scores (T and U) and loadings ($P^T$ and $Q^T$), E and F are residual matrix's equation 2.14 and 2.15.

$$X=TP^T+ E = \text{Structure + Noise} \qquad (2.14)$$

$$Y=UQ^T + F = \text{Structure + Noise} \qquad (2.15)$$

Scores (T) explain part of X which is related to Y and Score (U) explain part of Y which is related to X [12].

### 2.2.5 Variable selection

Variable selection is used for improving the model performance, give better predictions or reduce the model complexity by removing unnecessary, uninformative and interfering variables that add noise and makes prediction worse. Variable selection is a process of reducing number of independent variables in X matrix; by discriminating informative variables from the ones that are not related to dependent variable Y [25]. If too many variables are used the equation becomes over-fitted. This means the model will be data dependent and will give poor prediction results. On the contrary, using too few variables could result in under-fitting. This means the model is not large enough to capture the important variability in the data. Various variable selection approaches have been developed to reduce the complexity of the model [26]. A thorough understanding of data is necessary to make qualified decisions and get appropriate insight on what variable seems important, unimportant or is of intermediate importance. Chemical information from NIR spectra should be used when selecting variables to keep. Variable selection is an iterative process and should not be used as an automated black box selection approach.

### 2.2.6 Cross-Validation (CV)

Cross validation [27, 28] is a method used for evaluating predictive performance of a model. It is based on splitting the calibration data set into training set and test set, the process of splitting is repeated several times using different partition of the calibration data. The resulting prediction errors are averaged across the multiple rounds of CV.

In k-fold CV the data set is divided into k equal size subsets. Each time, one of the k subset is used as test set and k-1 subsets are used as training set to build model. The subset which was removed is then fitted to the model and the deviation between the actual response variable (y) and the predicted response variable (*y*ˆ) is used to obtain prediction error. The CV process is repeated k times, with each of the k subsamples used exactly once as the test data. The prediction error for all objects are then combined to obtain an overall prediction error given by root mean square error of cross validation (RMSECV) equation 2.16 [20] . This error is calculated for each number of LVs used to build the model. The number of LVs that archives lowest error is the optimal one.

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2.16)$$

Where n is number of objects (samples), $y_i$ actual response and $\hat{y}_i$ is predicted response

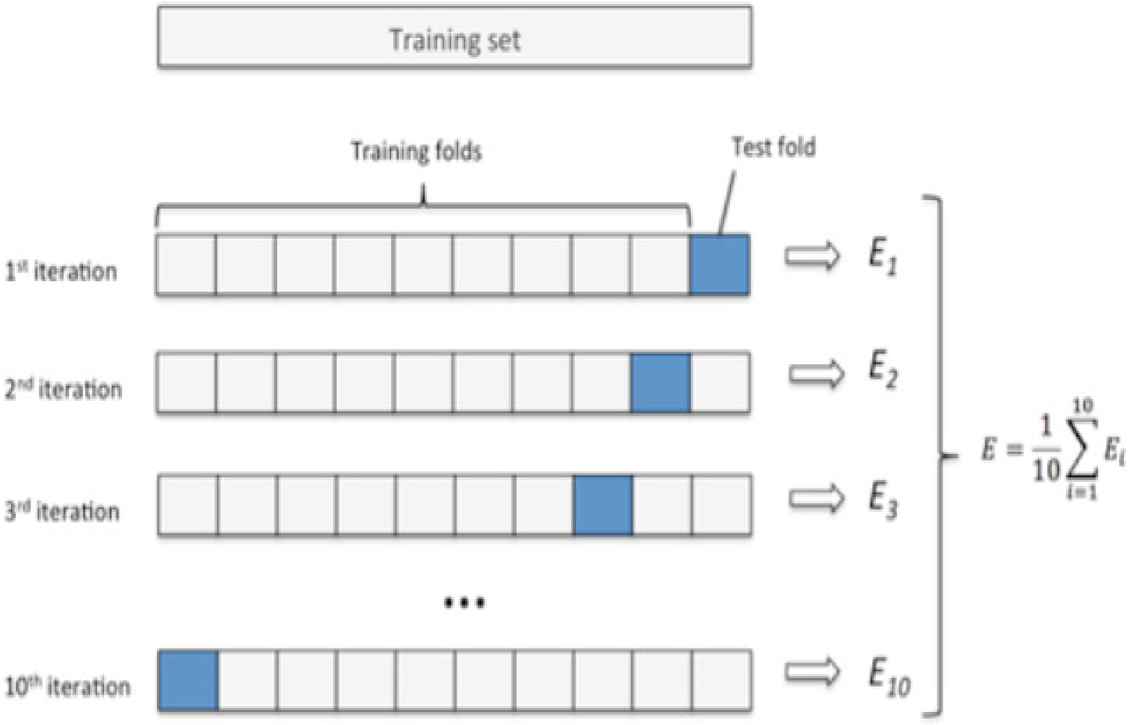


Figure 2.4: Representation of a 10 fold CV example. The calibration set is partitioned into 10 different groups. The error of each group (E1, E2 ….E10) is combined to obtain overall prediction error (E).

# 3 Experiment

## 3.1 Sample preparation

Thirty different commercial wheat gluten samples were collected from fish feed producing companies. Sample preparation for the experiment was done in three stages. Each of these stages are described in section 3.3 below.

## 3.2 Retsch Sample Divider PT 100 for sample splitting

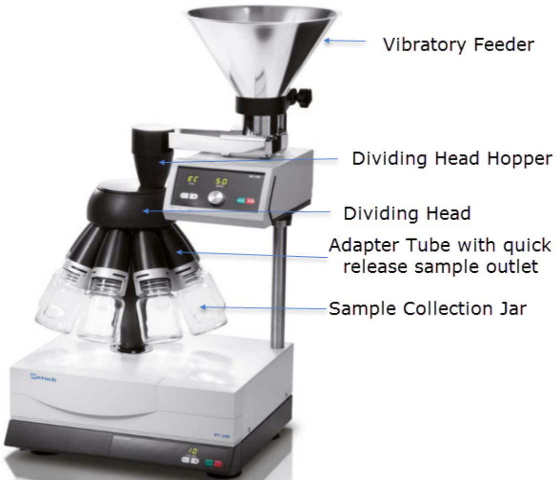


Figure 3.1: Retsch Sample Divider PT 100

Retsch Sample Divider PT 100 is used for splitting. It ensures representativeness of samples. In a retsch sample divider the material to be split flows through a vibratory feeder and is directed via a dividing head hopper into the opening of the dividing head. This dividing head is speed controlled and rotates at a constant speed of 110 revolutions per minute (rpm). The dividing head divides the sample evenly among the sample bottles that are attached to the adapter tube.

While splitting the wheat gluten samples it was observed that since wheat gluten is a very dry amorphous powder, a small amount of powder needs to be added to the vibratory feeder at a time. This was necessary to avoid blockages in the vibratory feeder.

## 3.3 Sample preparation and splitting

Sample preparation for the experiment was done in three stages, which are as described below.

**Stage one:** Obtaining representative sub samples

Stage one involved splitting each received wheat gluten sample into representative sub samples. This is done as follows.

1. Each of the 30 received wheat gluten samples (2.5 kg) was divided into 8 representative sub-samples (each being 280-290 grams approximately) by using a Retsch Sample divided PT100 as described under section 3.2.
2. Five sub-samples obtained were used to prepare test samples containing NPN compound at 5 different levels (500, 3000, 5500, 8000 and 10500 ppm), as described in stage 2 below.
3. One sub-sample was analyzed for microscopic analysis to ensure that the initial sample is pure and does not contain any impurities.
4. One sub-sample was used for pure wheat gluten scan on NIRS.
5. The remainder sample was stored as a backup sample.

The Flow chart figure 3.2 below show stage one splitting of the each of the 30 wheat gluten samples received.

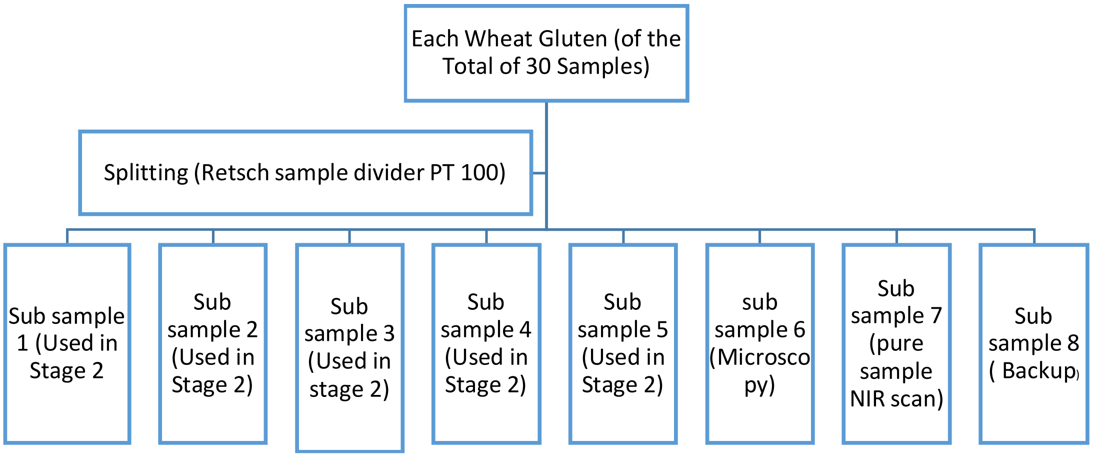

Each Wheat Gluten (of the Total of 30 Samples)
Splitting (Retsch sample divider PT 100)
Sub sample 1 (Used in Stage 2
Sub sample 2 (Used in Stage 2)
Sub sample 3 (Used in stage 2)
Sub sample 4 (Used in Stage 2)
Sub sample 5 (Used in Stage 2)
sub sample 6 (Microscopy)
Sub sample 7 (pure sample NIR scan)
Sub sample 8 ( Backup)

Figure 3.2: Flow chart to show stage one splitting of the each of the 30 wheat gluten samples received

**Stage two:** Mixing sub sample with NPN compound

Stage two was considered a very crucial step as it involved homogenous mixing of known amount of sub sample with the known amount of NPN compound to obtain test samples at desired concentration levels. Stage two consisted of the following steps:

1. Of the five sub samples obtained at stage one (from each wheat gluten sample), approximately 270- 285 grams of each sub sample was weighed using a weighing balance. Weight of the sub sample was noted down.

2. Required quantity of NPN compound to spike sub sample was weighed. Weight of NPN compound was noted down.

3. The weighed sub samples was mixed with weighed amount of NPN compound using mortar pestle to produce the desired concentration level as described below.

3.1 The sub sample with a weight of approximate 275-285 g was mixed with a NPN compound of 138-143 mg in order to attain a concentration level of 500 ppm

3.2 The sub sample with a weight of approximate 275-285 g was mixed with a NPN compound of 825-855 mg in order to attain a concentration level of 3000 ppm

3.3 The sub sample with a weight of approximate 275-285 g was mixed with a NPN compound of 1510-1575 mg in order to attain a concentration level of 5500 ppm

3.4 The sub sample with a weight of approximate 275-285 g was mixed with a NPN compound of 2200-2280 mg in order to attain a concentration level of 8000 ppm

3.5 The sub sample with a weight of approximate 275- 285 g was mixed with a NPN compound of 2890-3070 mg in order to attain a concentration level of 10500 ppm

4. Caution was followed to avoid samples spillage. The test samples obtained were collected in a self-sealing bag.

As a result, each wheat gluten sample was mixed with NPN compound to produce 5 different concentration levels. Thus, leading to 30 samples each, at concentration levels of 500 ppm, 3000 ppm, 5500 ppm, 8000 ppm and 10500 ppm. This generated a total of 150 samples at five different concentrations.

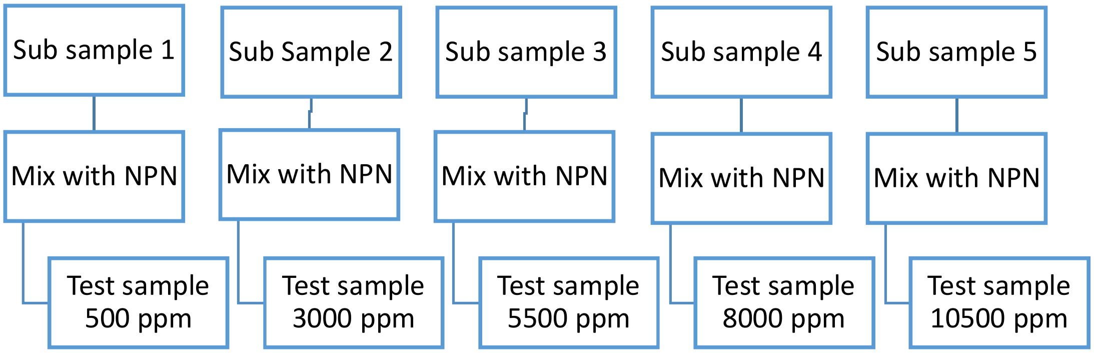


Figure 3.3: Flow chart to show 5 different levels of test sample, that were obtained from 5 sub samples of each wheat gluten sample.

**Stage three:** Splitting Test samples

In stage 3 each test sample obtained under stage 2 was split into three parts using Retsch sample divided PT100 as described under section 3.2. One of the split parts was used for NIR scanning, the second part was for reference analysis and the third part was stored as a backup sample.

The flow chart for the overall sample preparation and splitting is shown in figure 3.4

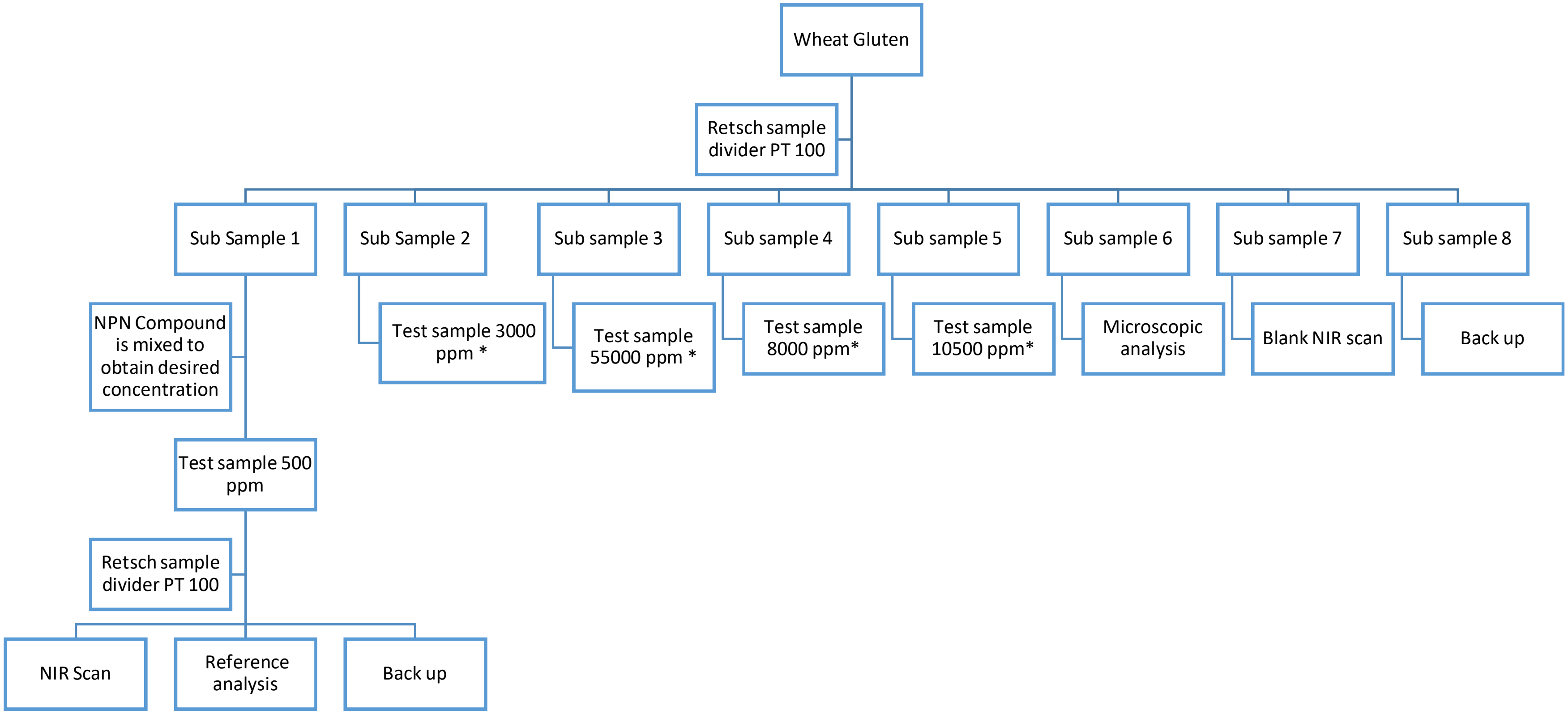
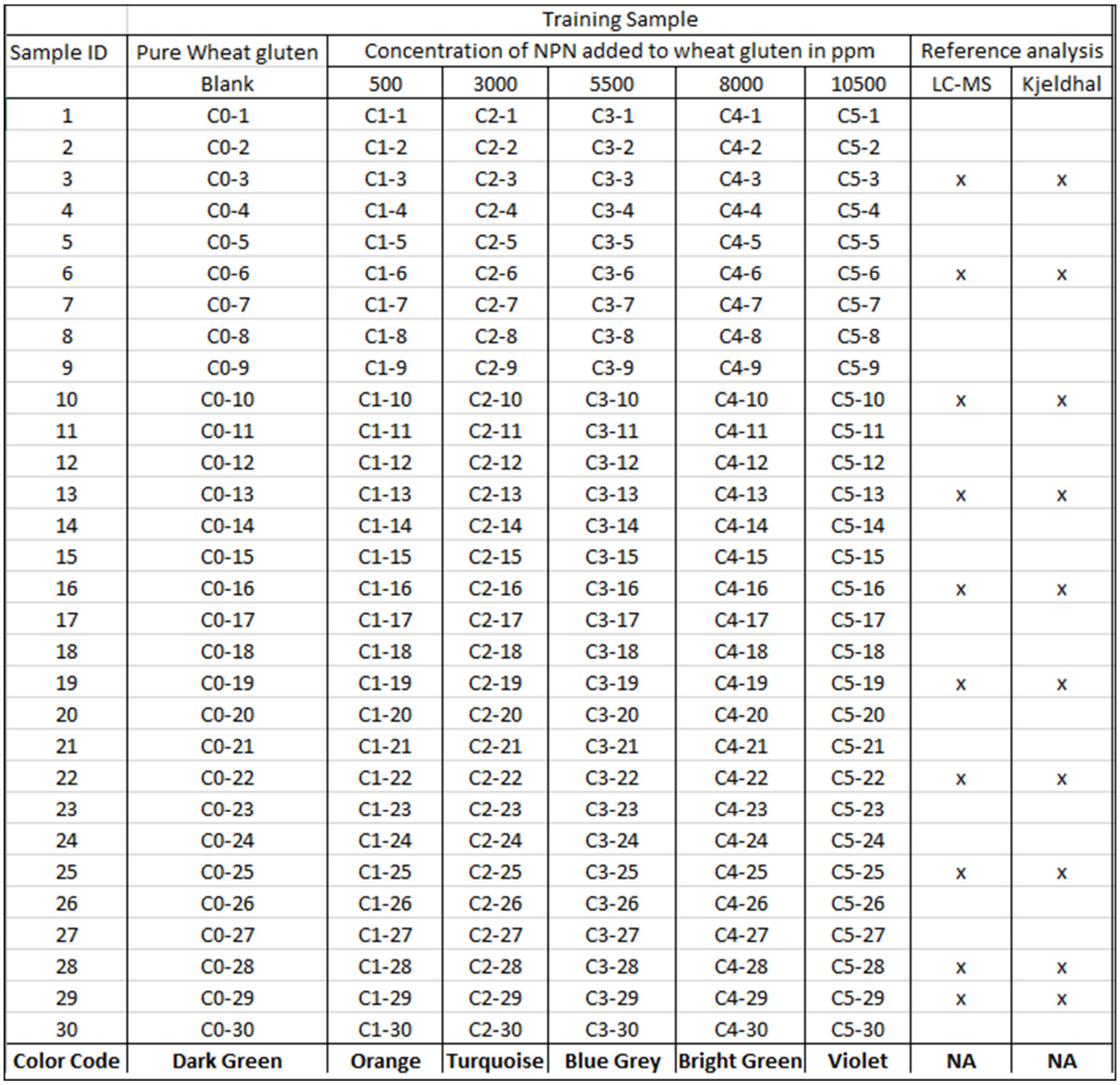


Figure 3.4: Flow chart representing overall sample preparation and splitting process of each wheat gluten sample

*Sub sample 2, 3, 4 and 5 have been treated in same way as sub sample 1 to obtain corresponding test samples

### 3.3.1 Sample marking

Table 3.1 gives information on the sample markings used to identify samples at different concentration, together with color code used to identify concentration levels during data analysis



| | Training Sample | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample ID | Pure Wheat gluten | Concentration of NPN added to wheat gluten in ppm | | | | | Reference analysis | |
| | Blank | 500 | 3000 | 5500 | 8000 | 10500 | LC-MS | Kjeldhal |
| 1 | C0-1 | C1-1 | C2-1 | C3-1 | C4-1 | C5-1 | | |
| 2 | C0-2 | C1-2 | C2-2 | C3-2 | C4-2 | C5-2 | | |
| 3 | C0-3 | C1-3 | C2-3 | C3-3 | C4-3 | C5-3 | x | x |
| 4 | C0-4 | C1-4 | C2-4 | C3-4 | C4-4 | C5-4 | | |
| 5 | C0-5 | C1-5 | C2-5 | C3-5 | C4-5 | C5-5 | | |
| 6 | C0-6 | C1-6 | C2-6 | C3-6 | C4-6 | C5-6 | x | x |
| 7 | C0-7 | C1-7 | C2-7 | C3-7 | C4-7 | C5-7 | | |
| 8 | C0-8 | C1-8 | C2-8 | C3-8 | C4-8 | C5-8 | | |
| 9 | C0-9 | C1-9 | C2-9 | C3-9 | C4-9 | C5-9 | | |
| 10 | C0-10 | C1-10 | C2-10 | C3-10 | C4-10 | C5-10 | x | x |
| 11 | C0-11 | C1-11 | C2-11 | C3-11 | C4-11 | C5-11 | | |
| 12 | C0-12 | C1-12 | C2-12 | C3-12 | C4-12 | C5-12 | | |
| 13 | C0-13 | C1-13 | C2-13 | C3-13 | C4-13 | C5-13 | x | x |
| 14 | C0-14 | C1-14 | C2-14 | C3-14 | C4-14 | C5-14 | | |
| 15 | C0-15 | C1-15 | C2-15 | C3-15 | C4-15 | C5-15 | | |
| 16 | C0-16 | C1-16 | C2-16 | C3-16 | C4-16 | C5-16 | x | x |
| 17 | C0-17 | C1-17 | C2-17 | C3-17 | C4-17 | C5-17 | | |
| 18 | C0-18 | C1-18 | C2-18 | C3-18 | C4-18 | C5-18 | | |
| 19 | C0-19 | C1-19 | C2-19 | C3-19 | C4-19 | C5-19 | x | x |
| 20 | C0-20 | C1-20 | C2-20 | C3-20 | C4-20 | C5-20 | | |
| 21 | C0-21 | C1-21 | C2-21 | C3-21 | C4-21 | C5-21 | | |
| 22 | C0-22 | C1-22 | C2-22 | C3-22 | C4-22 | C5-22 | x | x |
| 23 | C0-23 | C1-23 | C2-23 | C3-23 | C4-23 | C5-23 | | |
| 24 | C0-24 | C1-24 | C2-24 | C3-24 | C4-24 | C5-24 | | |
| 25 | C0-25 | C1-25 | C2-25 | C3-25 | C4-25 | C5-25 | x | x |
| 26 | C0-26 | C1-26 | C2-26 | C3-26 | C4-26 | C5-26 | | |
| 27 | C0-27 | C1-27 | C2-27 | C3-27 | C4-27 | C5-27 | | |
| 28 | C0-28 | C1-28 | C2-28 | C3-28 | C4-28 | C5-28 | x | x |
| 29 | C0-29 | C1-29 | C2-29 | C3-29 | C4-29 | C5-29 | x | x |
| 30 | C0-30 | C1-30 | C2-30 | C3-30 | C4-30 | C5-30 | | |
| **Color Code** | **Dark Green** | **Orange** | **Turquoise** | **Blue Grey** | **Bright Green** | **Violet** | **NA** | **NA** |

Table 3.1. An overview of the sample composition and color coding used.

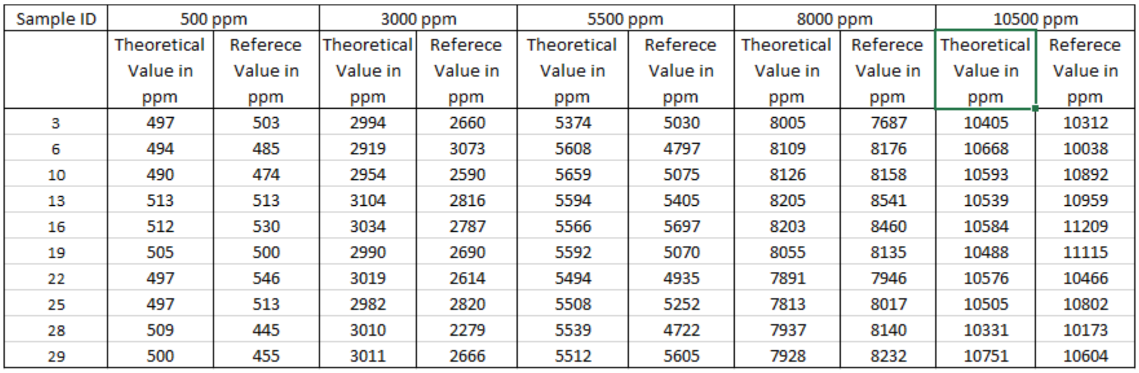
## 3.4 Reference method

### 3.4.1 Analysis of NPN Compound

Test samples obtained in stage 2 of sample preparation were prepared by mixing sub samples with NPN compound using mortar and pestle to ensure homogenous mixing of NPN compound. Ten test samples at each concentration level (total of 50-test sample) were sent to external lab for reference analysis of NPN compound using Liquid chromatography Mass spectroscopy (LC-MS) method. The results are given in table 3.2

The LC-MS method is validated by the external lab for lower concentrations of 0.1- 1.0 ppm and 1.0-100 ppm. The estimated measurement of uncertainty as provided by the external lab is 12.5%. This is based on extrapolation of validation data.

The samples marked as (x) in table 3.1, were analyzed to find concentration of NPN compound spiked . In the table 3.2 theoretical value is the value obtained by calculating known amount of NPN added to known amount of sub sample. Whereas concentration of NPN compound obtained by LC-MS method is marked as reference value. There is a good agreement between the theoretical value and the reference value. Thus, the method used for mixing NPN compound with the sub sample to prepare test sample was good enough to obtain homogenous samples.



| Sample ID | 500 ppm | | 3000 ppm | | 5500 ppm | | 8000 ppm | | 10500 ppm | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Theoretical Value in ppm | Referece Value in ppm | Theoretical Value in ppm | Referece Value in ppm | Theoretical Value in ppm | Referece Value in ppm | Theoretical Value in ppm | Referece Value in ppm | Theoretical Value in ppm | Referece Value in ppm |
| 3 | 497 | 503 | 2994 | 2660 | 5374 | 5030 | 8005 | 7687 | 10405 | 10312 |
| 6 | 494 | 485 | 2919 | 3073 | 5608 | 4797 | 8109 | 8176 | 10668 | 10038 |
| 10 | 490 | 474 | 2954 | 2590 | 5659 | 5075 | 8126 | 8158 | 10593 | 10892 |
| 13 | 513 | 513 | 3104 | 2816 | 5594 | 5405 | 8205 | 8541 | 10539 | 10959 |
| 16 | 512 | 530 | 3034 | 2787 | 5566 | 5697 | 8203 | 8460 | 10584 | 11209 |
| 19 | 505 | 500 | 2990 | 2690 | 5592 | 5070 | 8055 | 8135 | 10488 | 11115 |
| 22 | 497 | 546 | 3019 | 2614 | 5494 | 4935 | 7891 | 7946 | 10576 | 10466 |
| 25 | 497 | 513 | 2982 | 2820 | 5508 | 5252 | 7813 | 8017 | 10505 | 10802 |
| 28 | 509 | 445 | 3010 | 2279 | 5539 | 4722 | 7937 | 8140 | 10331 | 10173 |
| 29 | 500 | 455 | 3011 | 2666 | 5512 | 5605 | 7928 | 8232 | 10751 | 10604 |

Table 3.2: Gives comparison of theoretical value with actual value of NPN compound in the test samples.

### 3.4.2 Total nitrogen determination by Kjeldahl method

Ten pure wheat gluten and ten test samples at each concentration level (total 50 test samples) were analyzed by analytical lab using Kjeldahl method for protein determination. Kjeldahl analysis was done to check for the contribution of nitrogen by NPN compound. The results are given in table 3.3.

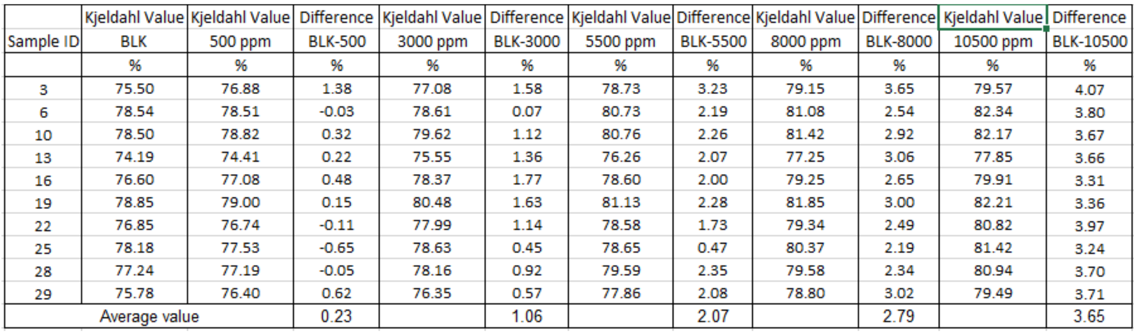
Kjeldahl [29] is a method for quantitative determination of total nitrogen content in substance. The Kjeldahl nitrogen determination method is made for the calculation of protein content in feeds, raw materials, forages and other samples. Kjeldahl method is recognized internationally for the estimation of protein content. It however does not give measure of true protein content as it measures non-protein nitrogen in addition to protein nitrogen in samples.

The procedure is carried out in three steps as follows.

1. Digestion: The sample is boiled in concentrated sulfuric acid and the nitrogen contained in the sample is converted to ammonium sulfate.
2. Distillation: Excess of sodium hydroxide solution is added to release ammonium ion in the form of ammonia, which is collected in the volumetric flask containing either boric acid, sulfuric acid or hydrochloric acid solution.
3. Titration: The amount of ammonia is then back titrated with sodium hydroxide solution.

The samples marked as (x) in table 3.1 were analyzed by Kjeldahl to determine nitrogen contribution of NPN compound at different spiked levels. In the table 3.3 Kjeldahl Value BLK is the kjeldahl nitrogen content of the pure wheat gluten sample. Kjeldahl value 500 ppm is nitrogen content of the test sample prepared to contain 500 ppm of NPN compound. Difference BLK-500 is the difference between two values obtained. Average value at the bottom of the table is the average difference between two readings.

The results from the Kjeldahl analysis show that as the level of spiking is increased the contribution of nitrogen by NPN compound is also increased. The average contribution of nitrogen at 500 ppm is 0.23% whereas the average contribution at 10500 ppm is 3.65%.Thus it can be concluded that, to generate profit from EMA, higher levels of NPN compound needs to be added.



| | Kjeldahl Value | Kjeldahl Value | Difference | Kjeldahl Value | Difference | Kjeldahl Value | Difference | Kjeldahl Value | Difference | Kjeldahl Value | Difference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample ID | BLK | 500 ppm | BLK-500 | 3000 ppm | BLK-3000 | 5500 ppm | BLK-5500 | 8000 ppm | BLK-8000 | 10500 ppm | BLK-10500 |
| | % | % | % | % | % | % | % | % | % | % | % |
| 3 | 75.50 | 76.88 | 1.38 | 77.08 | 1.58 | 78.73 | 3.23 | 79.15 | 3.65 | 79.57 | 4.07 |
| 6 | 78.54 | 78.51 | -0.03 | 78.61 | 0.07 | 80.73 | 2.19 | 81.08 | 2.54 | 82.34 | 3.80 |
| 10 | 78.50 | 78.82 | 0.32 | 79.62 | 1.12 | 80.76 | 2.26 | 81.42 | 2.92 | 82.17 | 3.67 |
| 13 | 74.19 | 74.41 | 0.22 | 75.55 | 1.36 | 76.26 | 2.07 | 77.25 | 3.06 | 77.85 | 3.66 |
| 16 | 76.60 | 77.08 | 0.48 | 78.37 | 1.77 | 78.60 | 2.00 | 79.25 | 2.65 | 79.91 | 3.31 |
| 19 | 78.85 | 79.00 | 0.15 | 80.48 | 1.63 | 81.13 | 2.28 | 81.85 | 3.00 | 82.21 | 3.36 |
| 22 | 76.85 | 76.74 | -0.11 | 77.99 | 1.14 | 78.58 | 1.73 | 79.34 | 2.49 | 80.82 | 3.97 |
| 25 | 78.18 | 77.53 | -0.65 | 78.63 | 0.45 | 78.65 | 0.47 | 80.37 | 2.19 | 81.42 | 3.24 |
| 28 | 77.24 | 77.19 | -0.05 | 78.16 | 0.92 | 79.59 | 2.35 | 79.58 | 2.34 | 80.94 | 3.70 |
| 29 | 75.78 | 76.40 | 0.62 | 76.35 | 0.57 | 77.86 | 2.08 | 78.80 | 3.02 | 79.49 | 3.71 |
| Average value | | | 0.23 | | 1.06 | | 2.07 | | 2.79 | | 3.65 |

Table 3.3: Test to check contribution of nitrogen by NPN compound at different spiked levels

# 4 Data analysis, results and discussion

## 4.1 Spectral acquisition

30 pure wheat gluten samples and 150 spiked samples (test samples) that were obtained after step 3 of sample preparation and splitting were analyzed using FOSS NIR instrument. Detail diagnostic test that gives information on the overall performance of the instrument was done before scanning test samples. It was ensured that the instrument was clean and all the samples were at room temperature. Each prepared NPN spiked sample (test sample) was scanned in duplicate between 400 nm- 2498 nm wavelength range with interval of 2 nm . Pure wheat gluten samples were scanned first followed by spiked test samples in the order of increasing concentration. This was done to prevent contamination of pure or lower concentration test samples with higher concentration test samples.

## 4.2 Software

Spectroscopic analysis was performed using FOSS NIR XDS Rapid Content™ Analyzer. NIR spectral data was collected using Foss ISIScan software version 4.10. The multivariate data analysis and modelling has been done using the program Sirius version 11.0 (Pattern Recognition System AS, Bergen, Norway software)

## 4.3 Multivariate Modelling

NIR spectra of pure wheat gluten figure 4.1 and of NPN compound figure 4.2 provide spectral signature rich in peaks. For the NIR spectra of NPN compound, three distinct peaks are seen around 1466, 1490 and 1520 nm. Cluster of peaks are also seen between 1974-2498 nm for both wheat gluten and NPN compound. The detection of contaminated samples was based on NIR spectra in the range from 1100-2498 nm region, as this region shows most of the peaks.

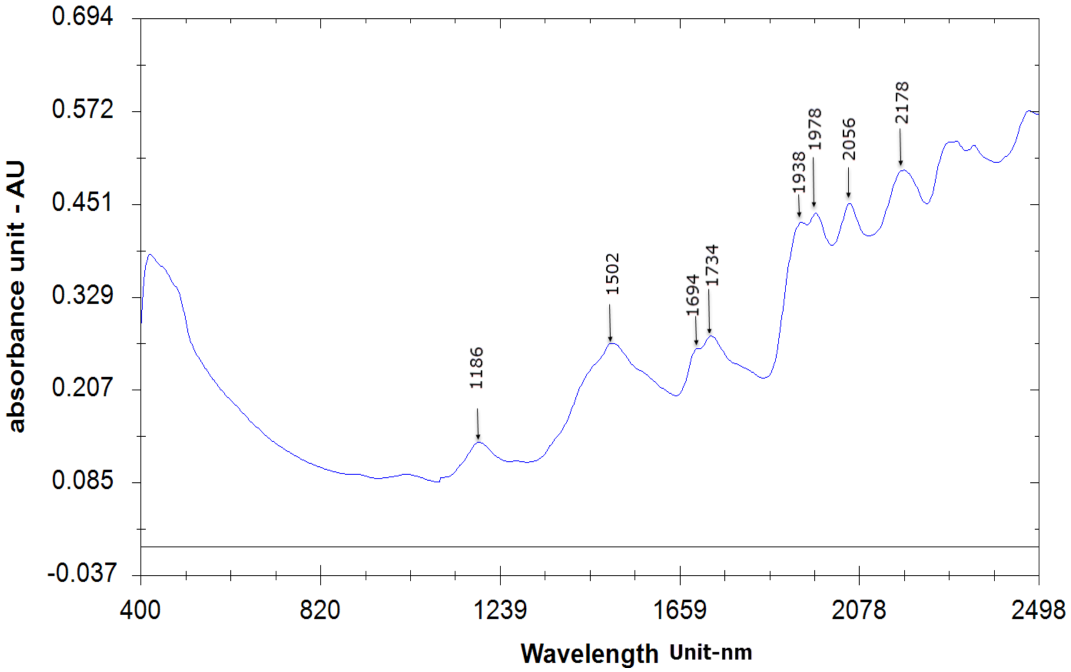


Figure 4.1: NIR Spectrum of pure wheat gluten recorded between 400- 2498 nm, showing major peaks

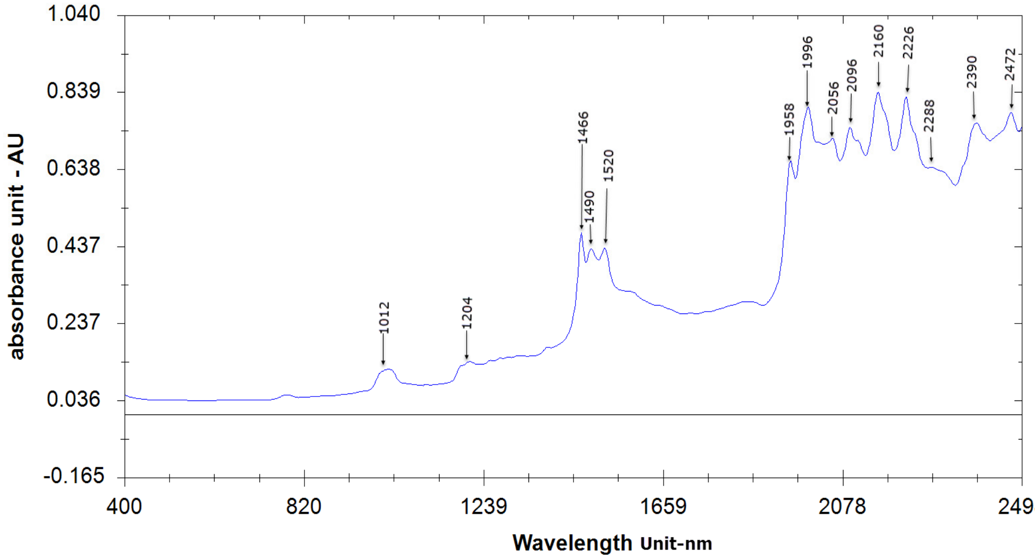


Figure 4.2: NIR spectra of NPN compound recorded between 400- 2498 nm, showing major peaks

NIR spectra often contain undesired scattering variation due to particle size, packing and sample surface amongst others. The scattering effect in NIRS consists of an additive effect and a multiplicative effect. The additive effect is reflected as baseline offset. The multiplicative effect is reflected as a slope that scales the entire spectrum. To minimize these complex baseline variations and scattering effects, data pre-treatment is needed. NIR spectra of the training set figure 4.3 was pre-processed using extended multiplicative scatter correction (EMSC) followed by second order differentiation width (nine), order (three) to eliminate these effects [19] . NIR spectra after scatter correction is shown in figure 4.4.

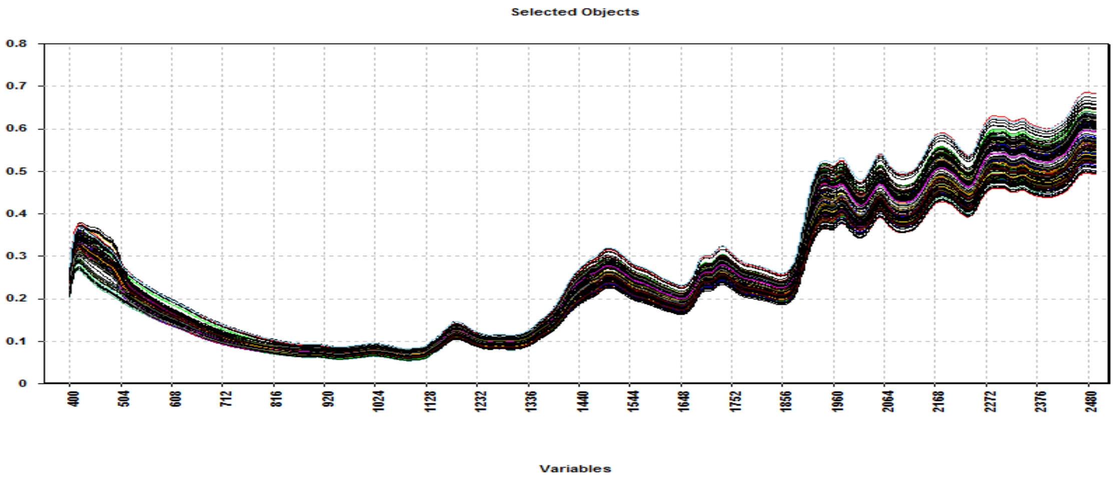


Figure 4.3: NIR Spectra of 30 pure wheat gluten samples and 150 test samples (30 at each concentration level (without any pre-processing)

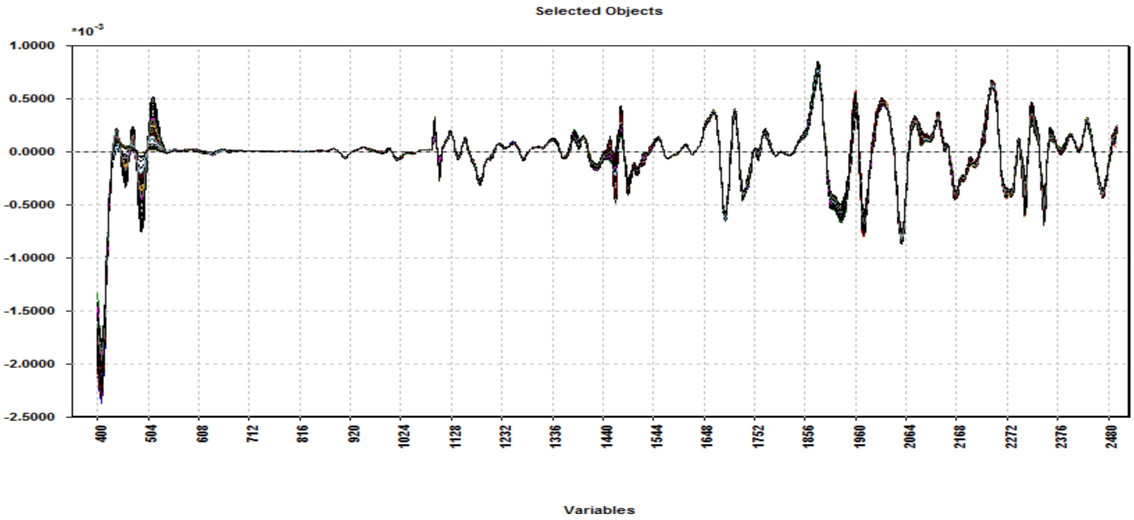


Figure 4.4: NIR Spectra of 30 pure wheat gluten samples and 150 test samples (30 at each concentration level) after applying extended multiplicative scatter correction together with second order derivative pre-processing method.

### 4.3.1 SIMCA Model

The main objective of this study is to develop SIMCA model that can differentiate between authentic and unauthentic wheat gluten samples. The SIMCA method builds individual PCA models, one for each class, and uses these to classify and discriminate new samples.  Cross validation was used to determine the number of significant components needed to describe the systematic variation in spectral data [30, 31] . NIR spectra was pretreated to compensate for scattering using EMSC followed by second order derivative width (nine) order (three). Spectral range where informative spectral difference between adulterated and non-adulterated wheat gluten samples was available, were selected to obtain an optimal calibration model. In this case spectral range used for best discrimination was 1100-2498 nm wavelength (refer figure 4.1 and 4.2). Performance of the developed SIMCA model was evaluated using the following criteria

1. PCA score plots.
2. The interclass distance between pure wheat gluten samples and the wheat gluten samples spiked with NPN compound
3. The acceptance or rejection rates of the samples used for the validation of the model.

Classification performance of SIMCA model was evaluated based on seventeen totally new pure wheat gluten samples and twenty totally new wheat gluten samples spiked with NPN compound at a concentration level of 15000 ppm, 20000 ppm, 30000 ppm and 35000 ppm (five samples at each concentration level) was used.

### 4.3.2 PLSR model for quantitative modeling

The additional objective of this study is to investigate the extent of adulteration which can be identified using PLSR model. Schematic diagram to represent PLSR model is given in figure 4.5. PLSR model was created with NIR spectra of the test samples (training set) in matrix X. The concentration of the NPN compound mixed to obtain these test samples was used as the reference value in the Y vector. The training set consisted of 30 pure wheat gluten samples and 150 spiked samples (preparation of the test samples is explained in detail in chapter 3.3).  In the current study PLSR model was validated by cross validation [32]. PLSR model was not validated using external validation set. This is because only 30 pure wheat gluten samples were available. These were not considered sufficient to make an independent and representative external validation set.

Due to the presence of numerous and correlated X variables there is a risk of “overfitting”, ie, a well fitted model with little or no predictive power. Hence, it is important to test predictive significance of each PLS component and stop when components start to be non-significant. The best PLS component selection was based on the following criteria [32].

1. Cross validation ratio ($C_{SV}SD$)

2. Explained variance in the dependent and independent variables

3. Lowest value of root means square error of cross validation

Cross validation ratio ($C_{SV}SD$) is a ratio of total prediction error of a model after including a new component, and the total residual standard deviation before this inclusion. If the ratio is less than one , new component is included in the model and the procedure continues with the calculation of next component [32, 33].

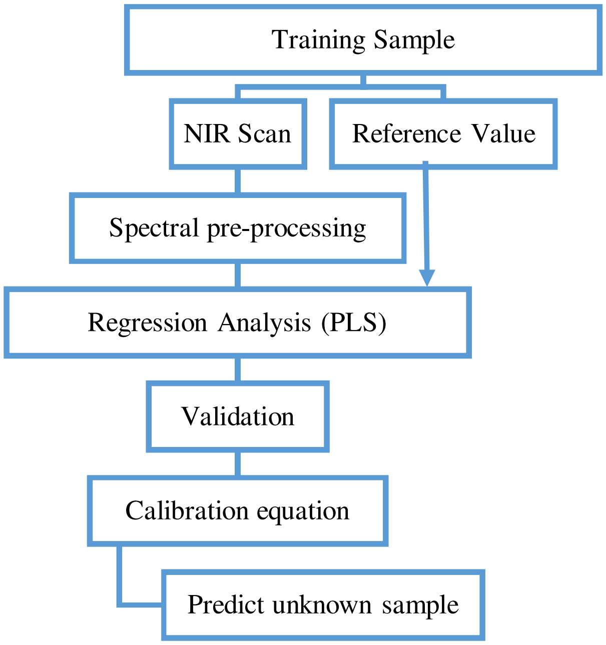


Figure 4.5: Schematic diagram to represent quantitative model

# 5 Results and discussion

## 5.1 SIMCA model

PCA was done on the whole training set (30 pure wheat gluten sample and 150 spiked wheat gluten samples, 30 at each concentration level) to look for groupings in the data. The data was mean centered and four PC's were extracted. The explained variance for four PC's is given in the table 5.1

| Principal Component | Explain variance |
|---|---|
| 1 | 58.71% (58.71%) |
| 2 | 27.93% (81.65% ) |
| 3 | 9.68% (91.33%) |
| 4 | 3.03% (94.36%) |

Table 5.1: Explained variance from the 4 PC's given by exploratory analysis of the whole training set

In the score plot different colors represent, different levels of spiking (refer table 3.1 for detail marking). It can be seen in the PC2 verses PC1 score plot figure 5.1 that the samples are not grouped based on the level of spiked NPN compound but are grouped based on the similarity between the wheat gluten samples. The figure 5.2 is score plot of PC 3 verses PC 1, it can be seen that samples are grouped on the bases of spiked level of NPN compound, but different groups overlap each other. The figure 5.3 is score plot of PC 3 verses PC 2. A better group separation is seen in this plot. Adding fourth PC does not seem to improve separation any further as seen in figure 5.4. Bar graph plot of scores verses objects for PC 3 is shown in figure 5.5. The bar graph shows scores of every object on the third PC. It can be seen that PC 3 to a larger extent explains the difference between different levels of spiking. Thus it can be concluded that three PC's are sufficient for separating samples based on spiked levels of NPN compound.

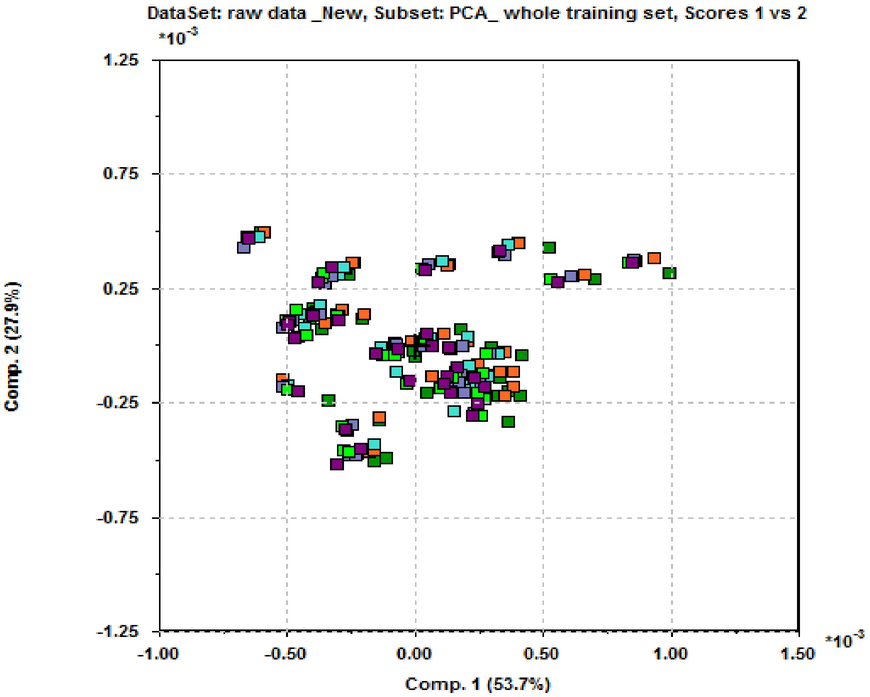


Figure 5.1: Score plot of PC2 verses PC1 (Different color represents different level of concentration)

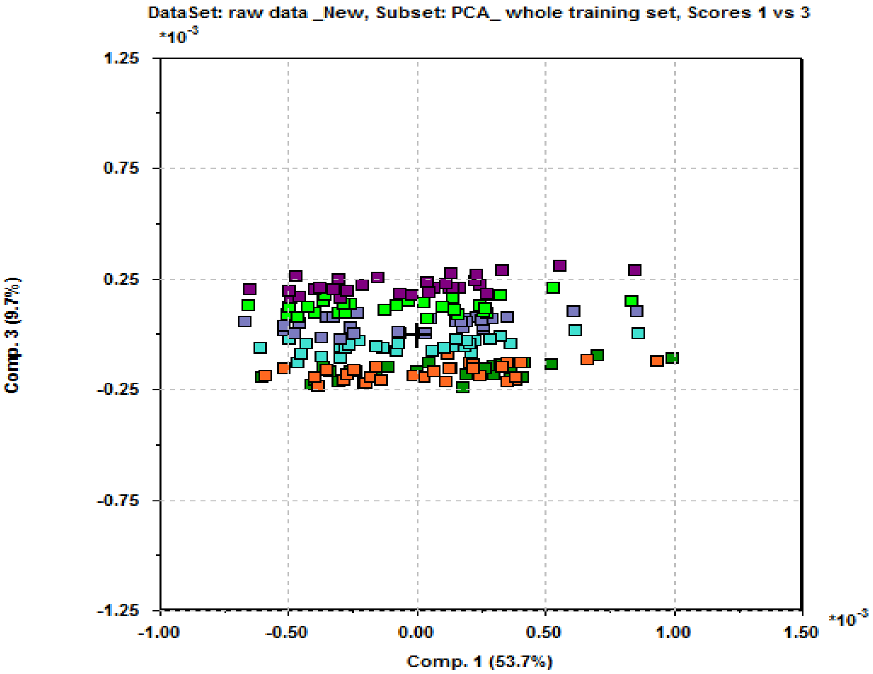


Figure 5.2: Score plot of PC3 verses PC1 (Different color represents different level of concentration)

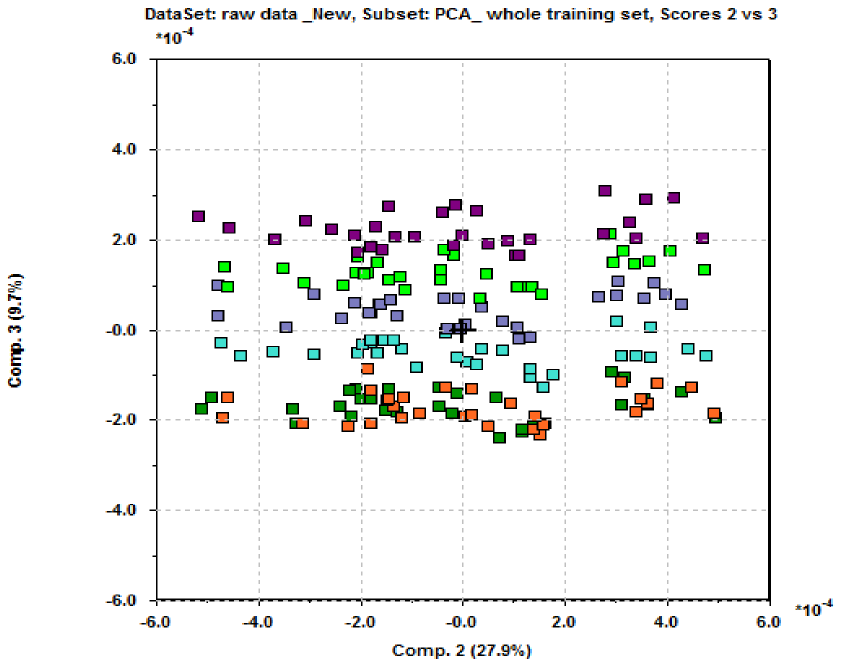


Figure 5.3: Score plot of PC3 verses PC2 (Different color represents different level of concentration)

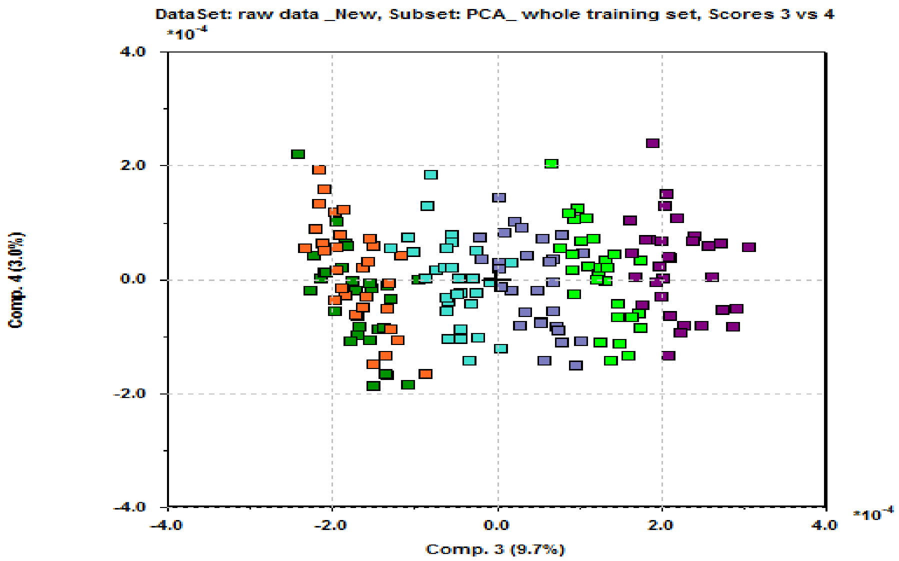


Figure 5.4: Score plot of PC4 verses PC3 (Different color represents different level of concentration)

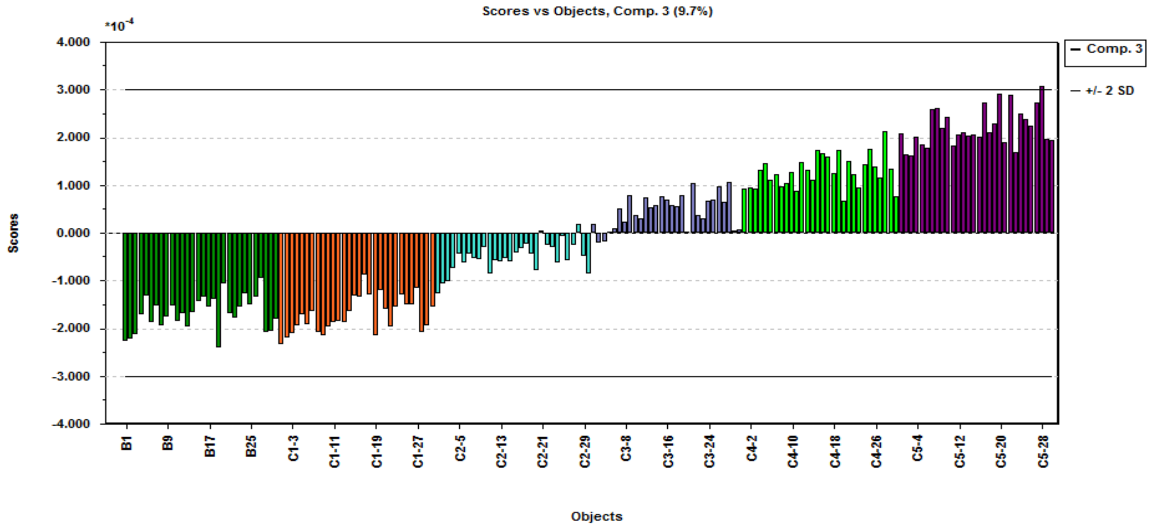


Figure 5.5: Bar graph of scores verses Objects

To create SIMCA model, separate PCA was performed for two different classes. One to include thirty pure wheat gluten samples and other to include thirty samples spiked with NPN compound at 10500-ppm concentration level. In the current study NPN compound was evaluated as EMA, meaning a higher concentration needs to be added to make adulteration economical viable. Adding 10500 ppm of NPN compound to the wheat gluten sample results in 3.7 % increase in protein levels (refer table 3.3). Hence the focus of this study is to create SIMCA model that can discriminate samples at 10500 ppm or higher.

Detailed information on the explained variance and cross validation ratio for PCA performed on thirty pure wheat gluten samples is given in table 5.2. As seen, PC 4 explains only 1.78% of the variance, which is very low, and has a high cross validation ratio of 0.93. Hence, a three PC model is considered.

| Principal Component | Explained variance | $C_{sv}SD$ |
|---|---|---|
| 1 | 58.98% (58.98%) | 0.72 |
| 2 | 30.76% (89.75% ) | 0.62 |
| 3 | 3.57% (93.32%) | 0.89 |
| 4 | 1.78% (95.10%) | 0.93 |

Table 5.2: Explained variance for pure wheat gluten class model

PCA model for the class with 30 wheat gluten samples spiked with NPN compound at 10500-ppm concentration was created. The results of explained variance and cross validation ratio is given in table 5.3. PC 4 explains only 1.49% of the variance in the data and has a high cross validation ratio of 0.92. Hence, a three-component model is considered for this subset.

| Principal Component | Explained variance | $C_{sv}SD$ |
|---|---|---|
| 1 | 58.85% (58.85%) | 0.73 |
| 2 | 32.36% (91.21%) | 0.58 |
| 3 | 3.5% (94.70%) | 0.88 |
| 4 | 1.49% (96.19%) | 0.92 |

Table 5.3 Explained variance for wheat gluten samples spiked with 10500-ppm of NPN

Modelling power for pure wheat gluten samples is given in figure 5.6 .Modelling power for spiked samples at 10500 ppm is given in figure 5.7. Discriminatory power of the two subsets is shown in figure 5.8. The two subsets have a discrimination power of 3.44. A distance greater than 3 indicates that the subsets are well separated and hence different [34].

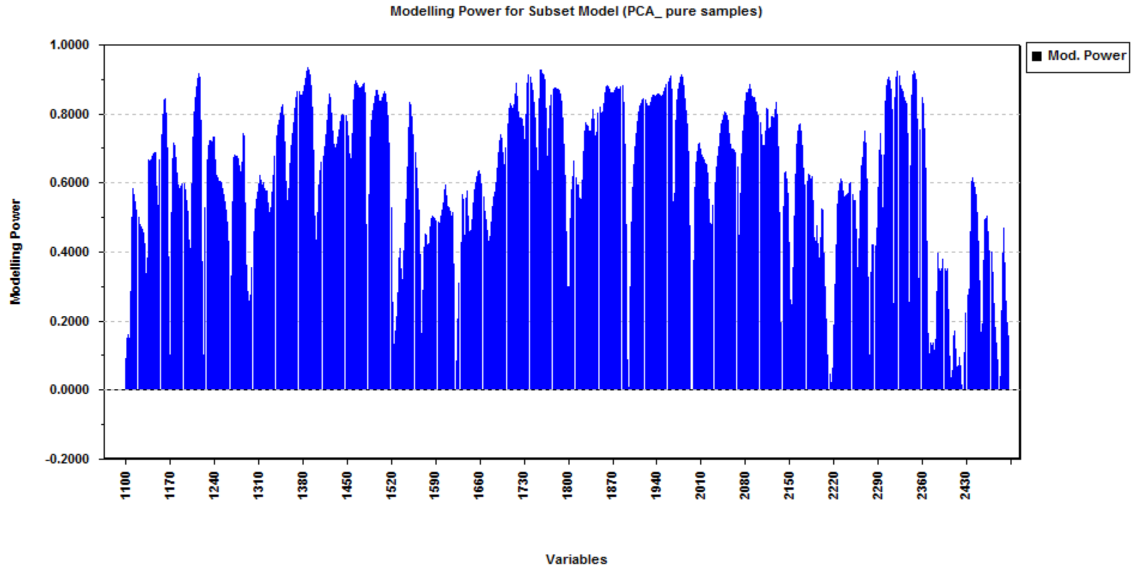


Figure 5.6: Modelling power plot of the pure wheat gluten sample

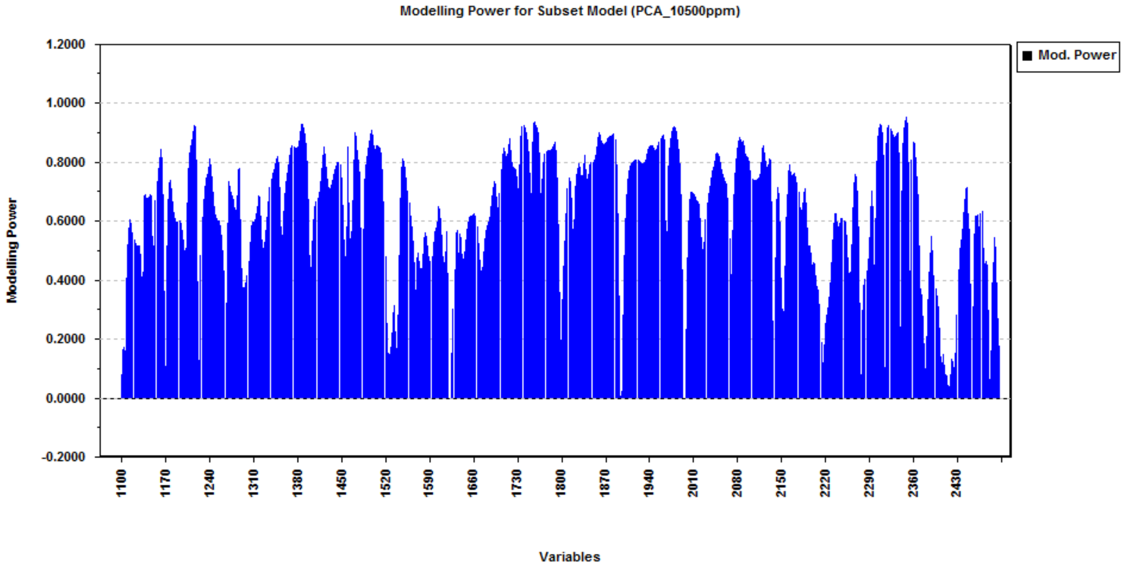


Figure 5.7: Modelling power plot of the spiked level 10500

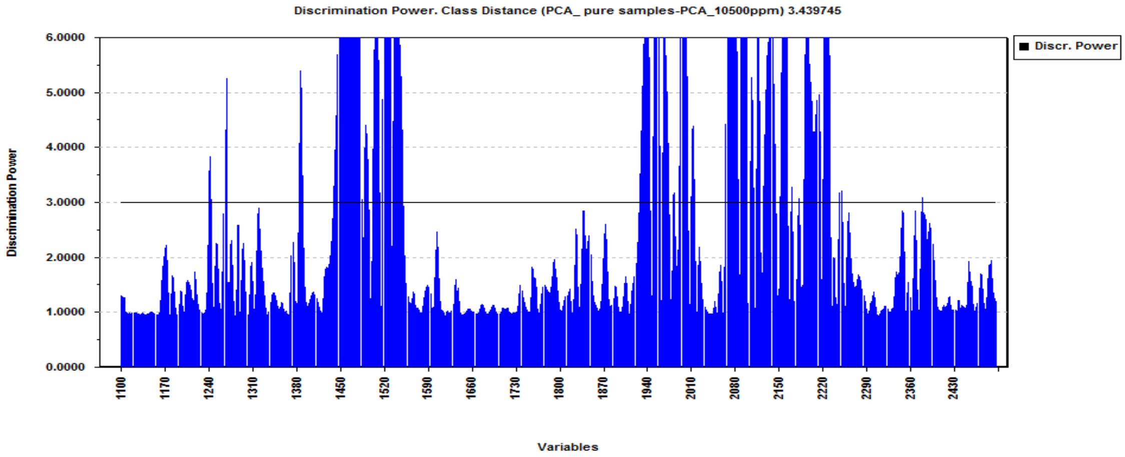


Figure 5.8: Discrimination power plot of the two-sub set created (pure wheat gluten and spiked samples at 10500 ppm)

Figure 5.9 is a plot of  RSD verses objects, of 30 pure wheat gluten and 150 spiked wheat gluten samples. It can be seen that as the concentration of the NPN compound is increasing, the spiked samples are moving away from the pure wheat gluten samples. There is overlap between pure wheat gluten samples and spiked samples at 500 and 3000 ppm but the samples at 5500 ppm (blue grey) and above have good separation from pure wheat gluten samples.

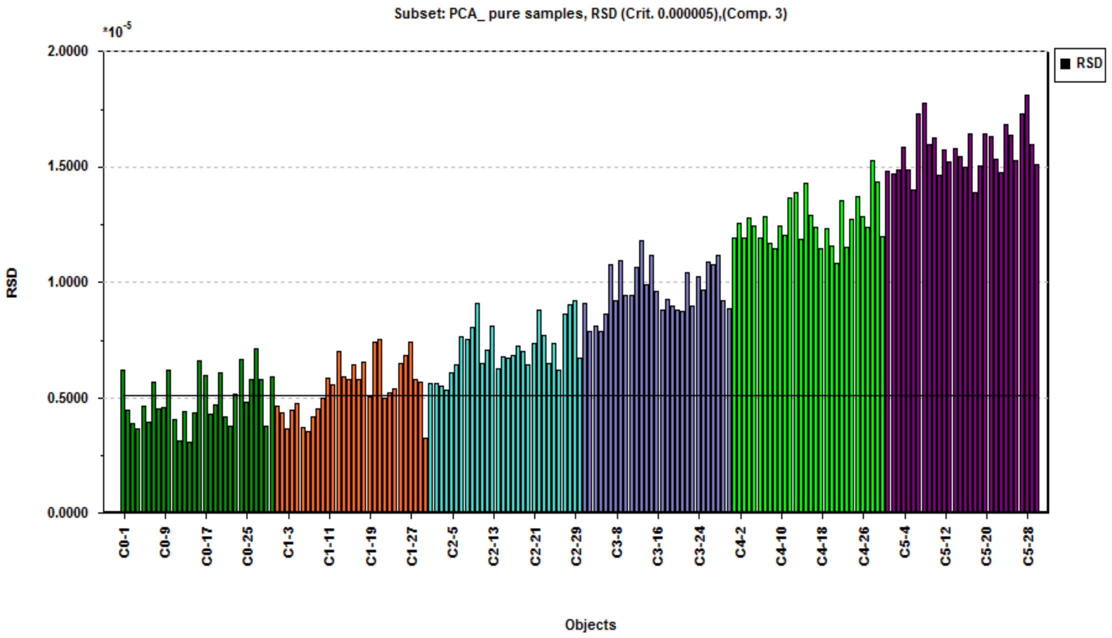


Figure 5.9: RSD verses object plot of pure wheat gluten and spiked wheat gluten sample at 5 different levels

(C0= pure wg , C1=500 ppm, C2= 3000 ppm, C3= 5500 ppm, C4= 8000 ppm and C5=10500 ppm)

**Evaluation of classification quality of SIMCA model using new sample set**

Classification quality of SIMCA model is associated with the expected errors of misclassification. Classification error are of two types: type I (false positive error) and type II (false negative error). The type I error occurs when an acceptable sample is classified as unacceptable during qualitative analysis. The type II error occurs when an unacceptable sample is classified as acceptable during qualitative analysis [35].

To evaluate classification quality NIR scan of seventeen totally new pure wheat gluten samples and twenty totally new wheat gluten samples spiked with NPN compound at a concentration level of 15000 ppm, 20000 ppm, 30000 ppm and 35000 ppm (five samples at each concentration level) was used.

The RSD value as shown in the figure 5.10 is very low. RSD value is calculated from F-test and has a very narrow confidence band due to strong correlation between the variables [30]. To resolve this problem the degree of freedom needs to be adjusted as each correlated variable does not contribute to a new degree of freedom [36].

An important factor for SIMCA classification is the number of PC's included in the model. It is a difficult task to determine the correct number of latent variables. For the current SIMCA model with 2 PC , 9 out of 17 pure new wheat gluten samples had RSD lower than the samples spiked at 10500 level. For SIMCA model with 3 PC, 10 out of 17 pure wheat gluten samples had RSD lower than the samples spiked at 10500 ppm level. However the overall RSD for 17 pure new wheat gluten samples was relatively low with 3 PC model. Hence a 3 PC model was used.

It is seen in the figure 5.10, that 10 out of 17 pure wheat gluten samples have RSD lower than 10500 ppm while 7 samples are misclassified as belonging to class with 10500 ppm or higher concentration spike levels. Hence there is a significant amount of type I error. This could be due to natural heterogeneity within the pure wheat gluten class, since the samples are coming from different batches, suppliers and different harvesting seasons etc. Including more samples will results in low type I error. The evaluation of type II error in such a classification system is very important. Type II error needs to be avoided as this type of error would present significant concern. This is done by subjecting the adulterated samples to the model, to check whether some of adulterated samples would be wrongly identified as belonging to the model. Of the twenty newly spiked samples none of the samples was wrongly identified as belonging to the pure wheat gluten class. This is an important result in the application of SIMCA for identification of adulterant in wheat gluten samples.

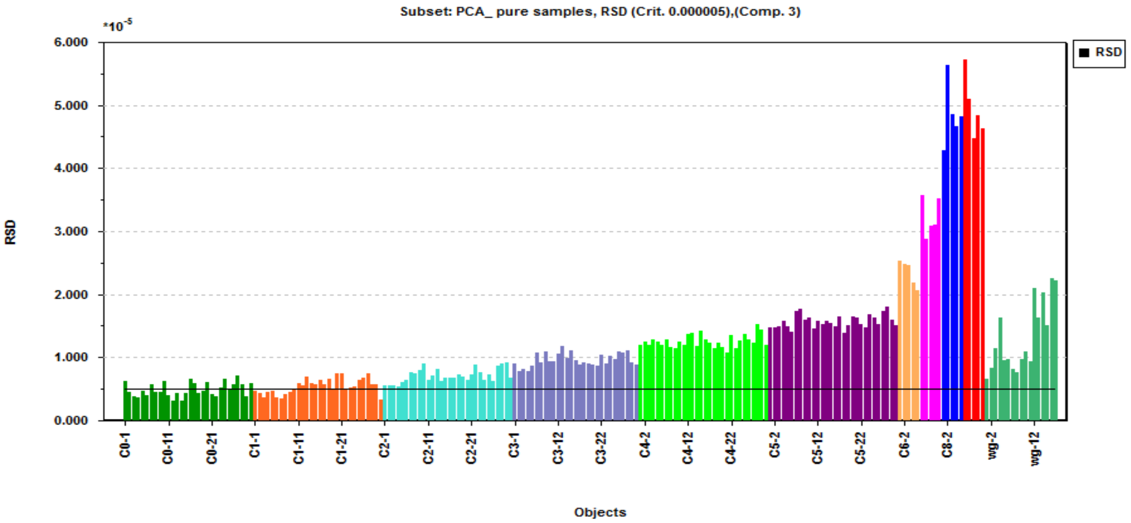


Figure 5.10: RSD verses objects (Where green= pure wheat gluten samples, orange = spiked at 500 ppm, turquoise = 3000ppm, blue gray = 5500 ppm, bright green = 8000 ppm, violet =

10500 ppm, light orange =15000, pink =20000 , blue =30000 , red =35000 , light green= 17 new pure wheat gluten samples )

**Enhancing model performance to improve classification errors**

Performance of the SIMCA model was improved by

1. Examination of the modelling and discriminating power plots, as it provides information on variables that are most important for separating the different classes.
2. Selecting variables related to chemical spectral information as seen in the figure 4.1 and 4.2.
3. Selecting variables with high discriminatory power for different spiked levels (3000, 5500 and 10500 ppm.

After variable selection exploratory analysis was done on the whole training set (30 pure wheat gluten samples 150 spiked samples) .The data was mean centered and three PC's were extracted . The explained variance for the three PC's is given in the table 5.4

| Principal Component | Explained variance |
|---|---|
| 1 | 70.48% (70.48%) |
| 2 | 21.49% (91.96%) |
| 3 | 5.99 % (97.96%) |

Table 5.4: Explained variance for three PC's, given by exploratory analysis after variable selection

As seen in the PC 2 verses PC 1 score plot figure 5.11 that the samples are grouped based on the level of spiked NPN compound. The figure 5.12 is score plot of PC 3 verses PC 1 here it can be seen that different groups form a cluster. Looking at the results from explorative data analysis, it can be concluded that two PC's are sufficient for separating samples based on spiked levels of NPN compound.

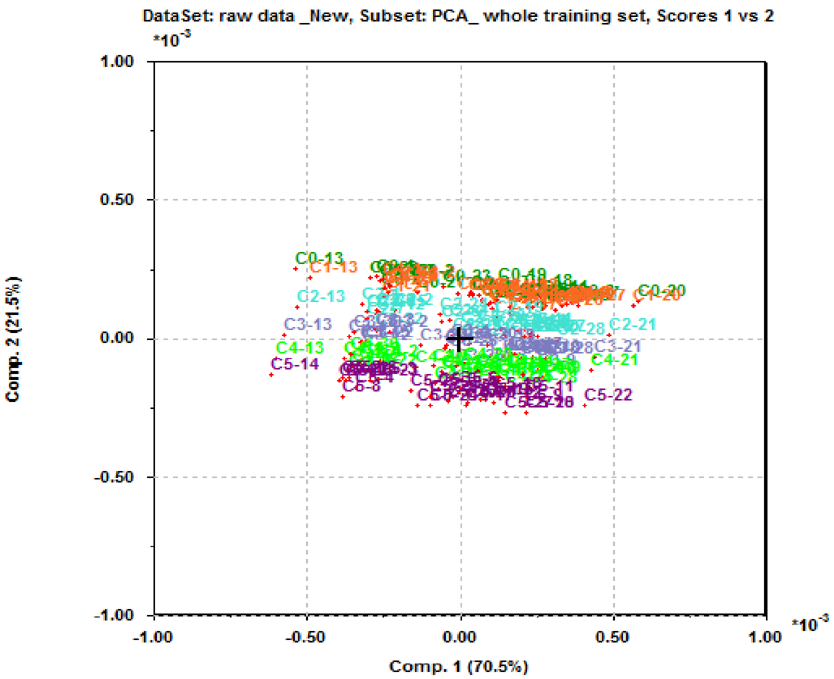


Figure 5:11: Score plot of PC 2 verses PC1 (Different color represents different level of concentration)

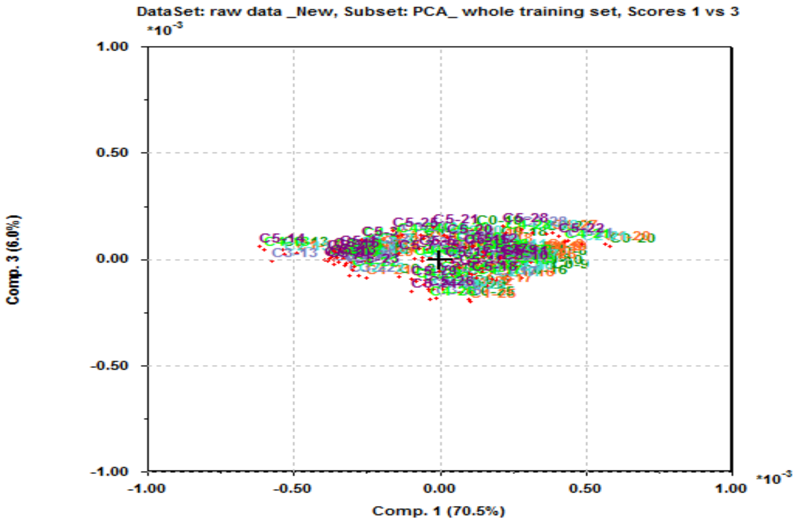


Figure 5:12: Score plot of PC 3 verses PC1 (Different color represents different level of concentration

Figure 5.13 shows the bar graph plot of score verses object. As can be seen in this figure, 3 PC 2 to a larger extend explains the difference between different levels of spiking. The variables that are selected to enhance SIMCA model performance are indicated by shading the area below the curve in figure 5.14

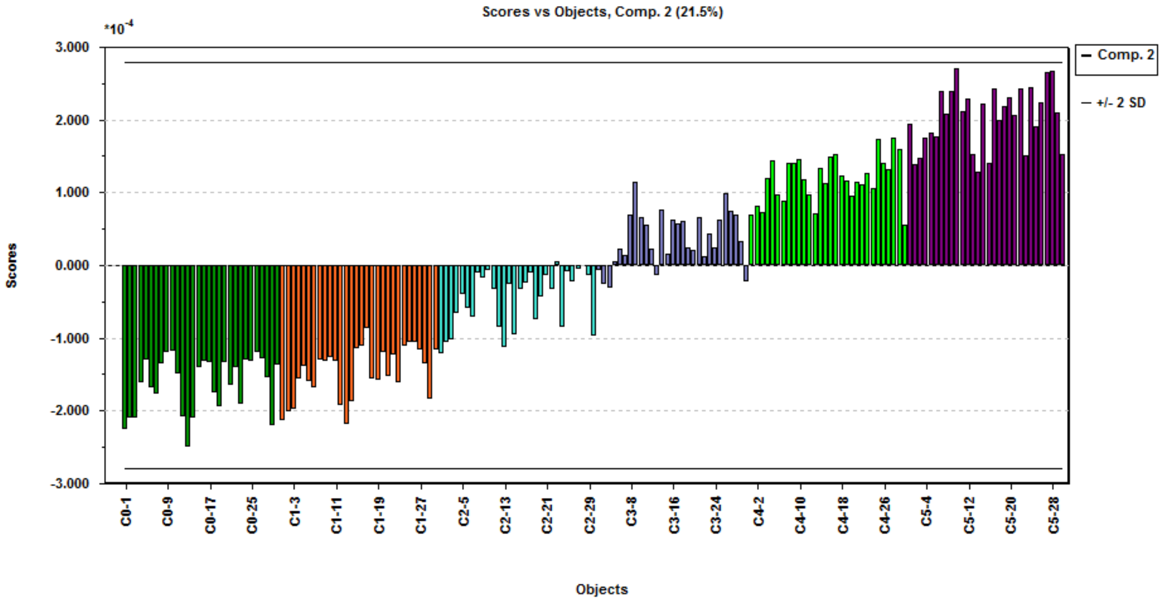


Figure 5.13: Bar graph of scores verses objects

After variable selection new PCA models were created for the pure wheat gluten samples and spiked samples at 10500 ppm. Explained variance and cross validation ratio for the pure wheat gluten class is shown in table 5.5. Explained variance after second PC is 97.33% with the cross validation ratio of 0.68. The cross validation ratio for the third PC is 0.92 and it explains only 0.72% of the variance. Thus, two PC are sufficient to explain the total variance in the model.

| Principal Component | Explained variance | $C_{sv}SD$ |
|---|---|---|
| 1 | 89.50% (89.50%) | 0.48 |
| 2 | 7.83% (97.33%) | 0.68 |
| 3 | 0.72% (98.05%) | 0.92 |

Table 5.5: Explained variance for new pure wheat gluten model after variable selection

Explained variance and cross validation ratio for the wheat gluten samples spiked at 10500-ppm is shown in table 5.6. Explained variance after third PC is 98.10% with the cross validation ratio of 0.85. The cross validation ratio for the fourth PC is 0.9 and it explains only 0.62% of the variance. Thus, three PC's are good enough to explain the total variance in the model.

| Principal Component | Explained variance | $C_{sv}SD$ |
|---|---|---|
| 1 | 88.37% (88.37% ) | 0.5 |
| 2 | 8.56% (96.93%) | 0.63 |
| 3 | 1.16% (98.10 %) | 0.85 |
| 4 | 0.62% (98.71%) | 0.9 |

Table 5.6: Explained variance for 10500 ppm samples model after variable selection

The discriminatory power for the two subsets (pure wheat gluten and sample spiked at 10500 ppm) is shown in figure 5.14. As can be seen, after variable selection discriminatory power was increased to 7.50 compare to 3.44 without variable selection. As seen in NIR scan from NPN compound (figure 4.2) three distinct peaks were seen around 1466, 1490 and 1520 nm that seem to be important to discriminate between adulterated wheat gluten with pure wheat gluten. These wavelengths have higher discriminatory power.

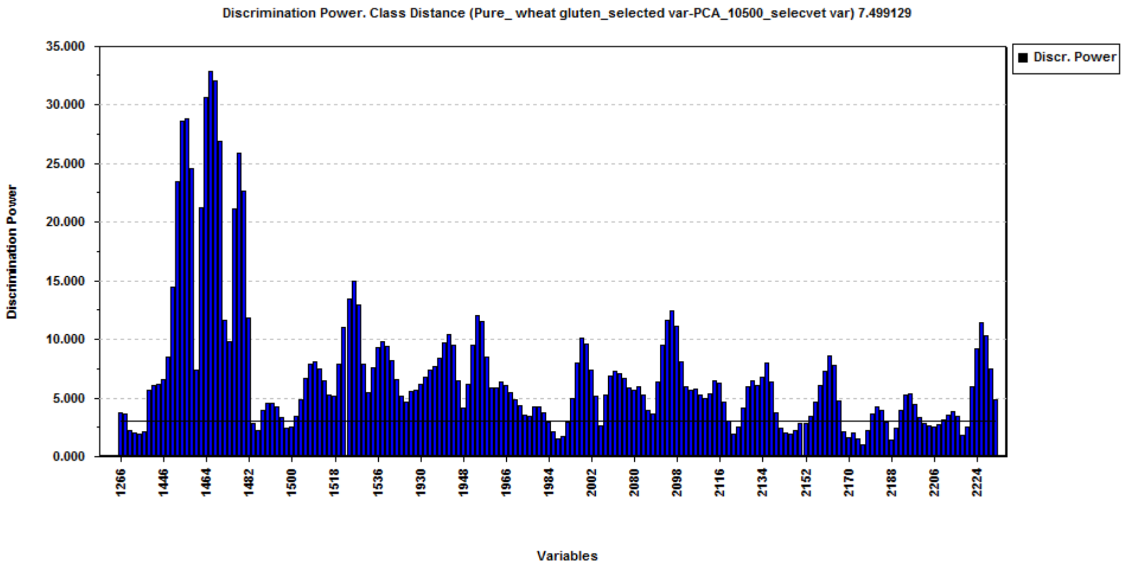


Figure 5.14: Discrimination power plot after variable selection

Figure 5.15 shows a plot of RDS verses object for 30 pure wheat gluten and 150 spiked wheat gluten samples. It can be seen that , with the redefined SIMCA model it is possible to differentiate sample at spike level of as low as 3000 ppm with 100 % classification rate. The pure wheat gluten sample and samples at 500 ppm spike levels still show some overlap.

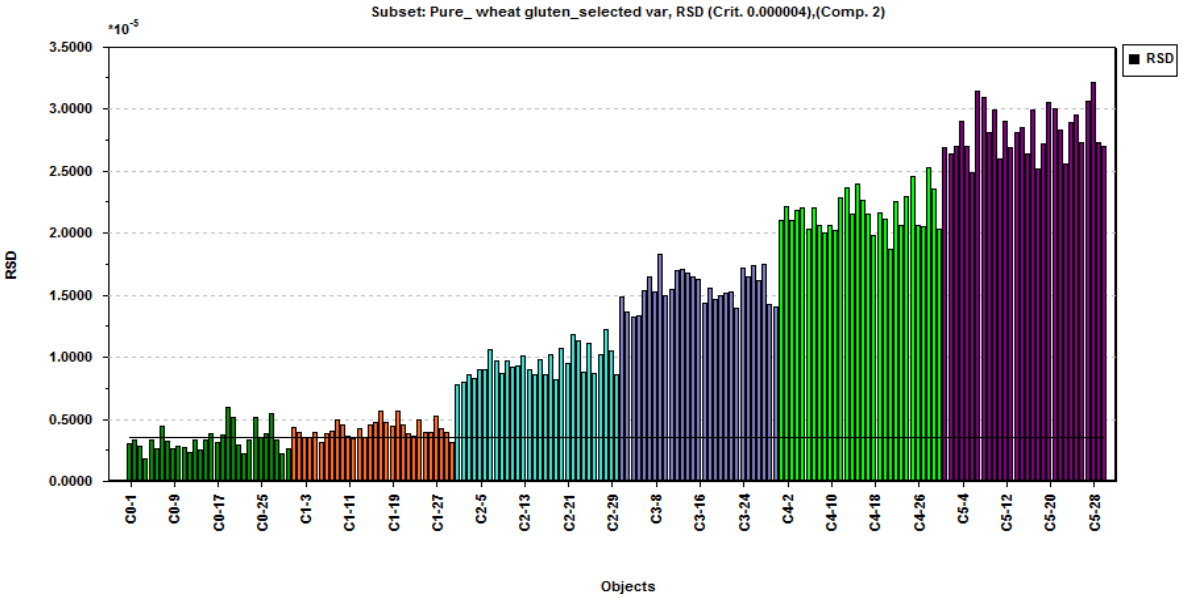


Figure 5.15: RSD verses object plot of pure wheat gluten and spiked wheat gluten sample at 5 different levels with redefied SIMCA model

(C0= pure wg , C1=500 ppm, C2= 3000 ppm, C3= 5500 ppm, C4= 8000 ppm and C5=10500 ppm)

### Evaluation of new SIMCA model created after variable selection

Classification performance of the SIMCA model created after variable selection was evaluated using NIR scan of seventeen new pure wheat gluten samples and twenty new spiked wheat gluten samples at concentration level of 15000 ppm, 20000 ppm , 30000 ppm and 35000 ppm (five sample at each concentration level). Figure 5.16 show plot of RSD verses object. It can be seen in the figure that the twenty new spiked samples standout as not belonging to the pure wheat gluten class samples. The RSD for seventeen pure new wheat gluten samples is close to the RSD for samples spiked with NPN at 500 ppm concentration. Thus the misclassification rate is zero for type I and type II error. These results show that SIMCA model could be developed to discriminate the spectral signals of adulterated and non-adulterated wheat gluten samples at a level as low as 3000 ppm with 100 % classification.

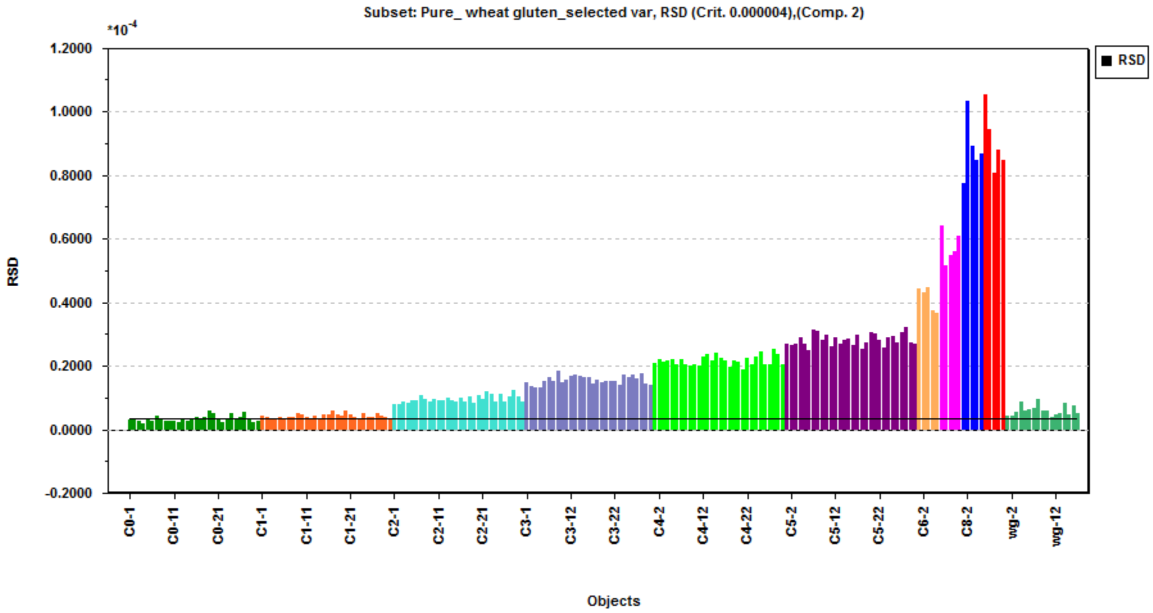


Figure 5.16: Plot of RSD verses object.

(Where green= pure wheat gluten samples, orange = spiked at 500 ppm, turquoise = 3000ppm, blue gray = 5500 ppm, bright green = 8000 ppm, violet = 10500 ppm, light orange =15000ppm, pink =20000ppm, blue =30000ppm , red =35000ppm , light green= 17 new pure wheat gluten samples )

## 5.2 PLS Model

The process of deriving best PLS equation was carried out in the following steps

**Step 1:** Cross validation results of initial PLSR analysis showed that five PLS components gave the best prediction performance figure 5.17. Detailed information on the independent and dependent variables together with cross validation ratio is given in table 5.7. The first two components explain 97.05% of the variance in Y. When the third component is added explained variance in Y is increased only by 0.41%. The third component explains very little of the variance. The second component has a $C_{SV}$SD value of 0.37 and it increases to 0.88 for the third component. Lower value for the cross validation ratio is preferred. Thus including more than two PLS components could lead to overfitting. This indicates that a five component PLS model is not optimum for predictions.

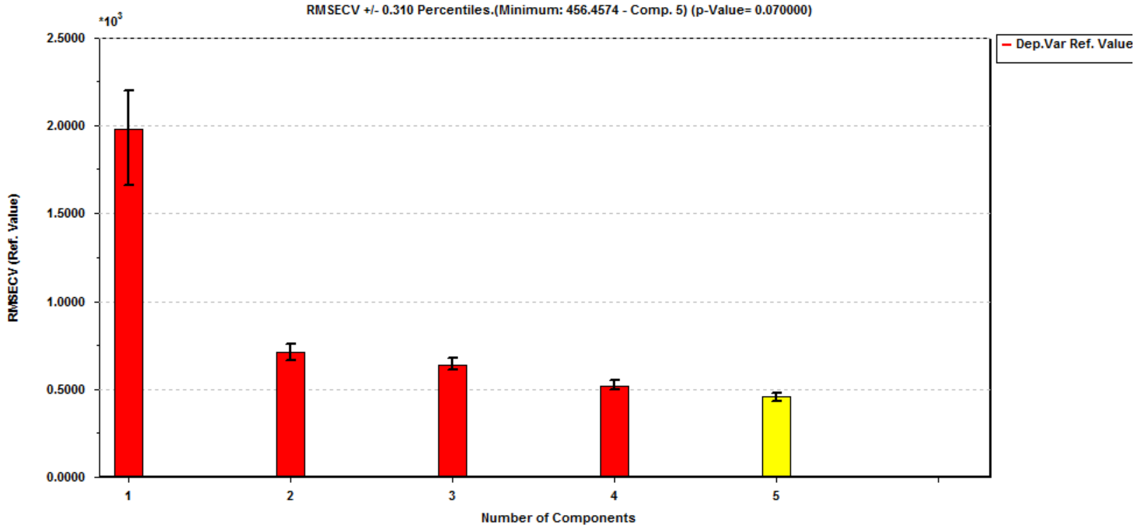


Figure 5.17: Plot of RMSECV value with number of components

| Component | Explained variance in independent | Explained variance in dependent | $C_{sv}SD$ |
|---|---|---|---|
| 1 | 23.59% (23.59%) | 76.60% (76.60%) | 0.5 |
| 2 | 39.77% (63.36%) | 20.45% (97.05%) | 0.37 |
| 3 | 24.28% (87.64%) | 0.41% (97.45%) | 0.88 |
| 4 | 6.48% (94.12%) | 0.90% (98.35%) | 0.83 |
| 5 | 1.16% (95.28%) | 0.49% (98.83%) | 0.85 |

Table 5.7: Explained variance for independent and dependent variable for 5 PLS components

**Step 2:** Regression coefficient analysis was performed to identify less important variables.

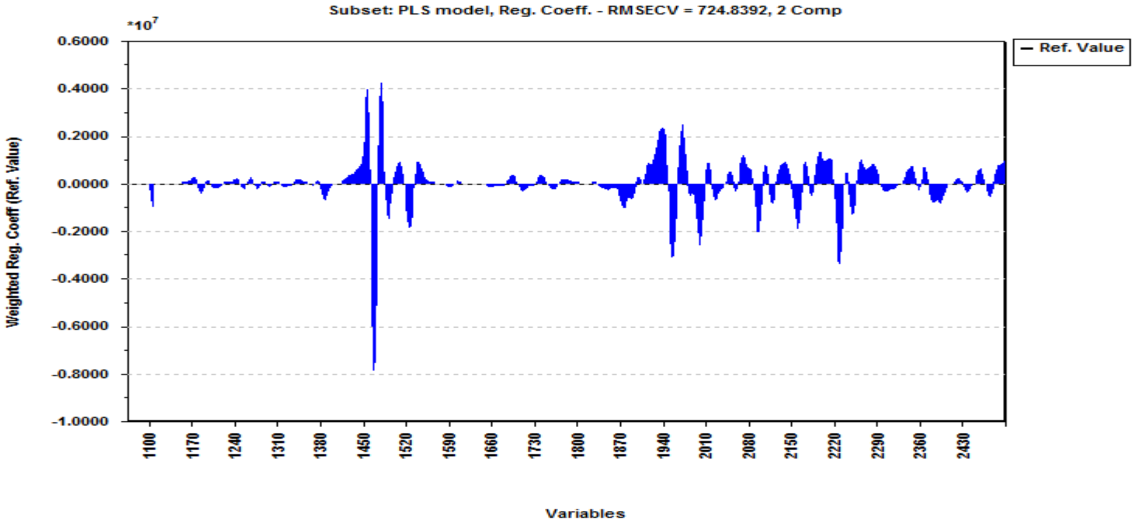


Figure 5.18: Regression coefficient plot

As seen in the figure 5.18, the region between 1100-1400nm and 1600-1900 nm is of little interest. In addition, similar region were also seen as less important in the NIR spectra of pure wheat gluten and NPN compound (refer figure 4.1 and 4.2). By omitting these variables, new PLSR model was created. The explained variance in the independent and dependent variable along with $C_{SV}$SD value is given in the table 5.8

| Component | Explained variance in independent | Explained variance in dependent | $C_{sv}$SD |
|---|---|---|---|
| 1 | 28.38% (28.38%) | 74.81% (74.81%) | 0.5 |
| 2 | 39.62% (68.00%) | 22.82% (97.62%) | 0.34 |
| 3 | 13.34% (81.35%) | 0.50% (98.12%) | 0.92 |
| 4 | 13.46% (94.81%) | 0.40% 98.52%) | 0.84 |

Table 5.8: Explained variance for independent and dependent variable for four component PLS model after omitting variables of little interest.

As seen from table 5.8 the first two components explain 97.63% of the variance in Y. When the third component is added explained variance in Y is increased only by 0.50%. The second component has a $C_{SV}$SD value of 0.34 and it is increased to 0.89 for the third component. The RMSECV value for the two component PLS model has increased to 637 compared to 456 for five component PLS model in step 1. This indicates that the model needs to be further refined, to find the most relevant variables and to obtain a model with a better predictive ability and lower RMSECV value.

**Step 3:** Selectivity ratio [26, 37] was used as a method for variable selection. Selectivity ratio is defined as the ratio between the explained variance of each variable to the residual variance. Figure 5.19 shows a plot of variable selectivity ratio verses variables .The selectivity ration can be displayed similarly to a spectrum. A small ration increases the risk of selecting false variables, while a high ration increases risk of losing important variables.

For the selection of variables based on selectivity ratio, following was also used as a guide.

1. Information from preliminary regression co-efficient analysis figure 5.18.
2. Variables with high discrimination power as seen in SIMCA model.
3. Selecting variables related to chemical spectral information as seen in figure 4.1 and 4.2.

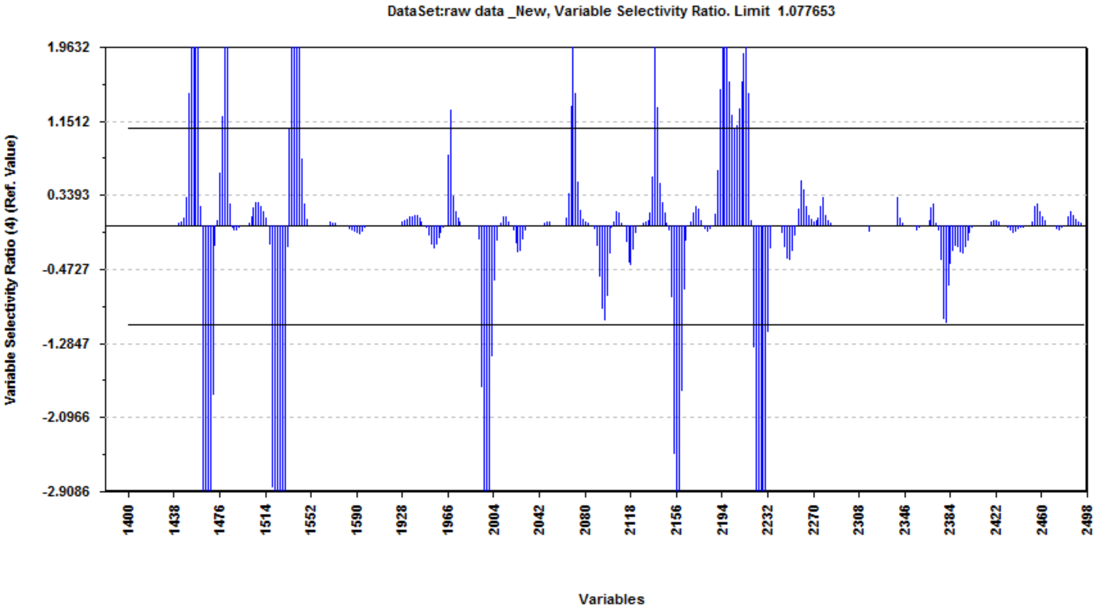


Figure 5.19: Selectivity ratio graph

As seen in the graph figure 5.19 above selectivity ratio suggested by program is 1.07. Most of the variables that have high discrimination power for the SIMCA model are included at this level. The distinct peaks seen in the figure 4.1 and 4.2 are also included.

After variable selection by selectivity ratio, new PLSR model was created. Table 5.9 below gives detail information on the explained variance and $C_{SV}SD$ . The total explained variance in dependent variable by first two component is 98.68 % and adding third component will increase it only by 0.22% .The $C_{SV}SD$ for the second component is 0.49 and is increased to 0.95 for the third component. In addition, the total explained variance in independent variable with first two component has increased to 95.49 % compared to 63.36 % and 68 % in step 1 and step 2 respectively (refer table (5.7) and (5.8)). The RMSECV value for two component PLS model is 459

Thus, a two-component PLSR model based on selected variables as described in step 3 was created.

| Component | Explained variance in independent | Explained variance in dependent | $C_{sv}SD$ |
|---|---|---|---|
| 1 | 82.29% (82.29%) | 93.97% (93.97%) | 0.24 |
| 2 | 12.85% (95.14 %) | 4.72% (98.69%) | 0.47 |
| 3 | 1.72% (96.86%) | 0.20% (98.88%) | 0.95 |
| 4 | 1.22% (98.09%) | 0.12% (99.01%) | 0.98 |
| 5 | 0.51% (98.60%) | 0.05% (99.06%) | 0.98 |

Table 5.9: Explained variance for independent and dependent variables for PLS model after variable selection using selectivity ratio.

The measured value of NPN compound for samples in the training set were plotted against predicted values of NPN compound (figure 5.20) using the two component PLS model created in step 3 .The correlation here is 0.993. This indicates a good fit of the model to the training set data. The objects are grouped in six clusters. This indicates that the dataset consists of six groups of objects. It is seen that two clusters that lie to the left, dark green (pure wheat gluten samples) and orange (500 ppm spiked test sample) are very close. Thus, the model might not be able to accurately predict samples with lower concentration levels of NPN compounds (500 ppm and lower). Overlap between pure wheat gluten and spiked samples at 500 ppm was also observed for SIMCA model.

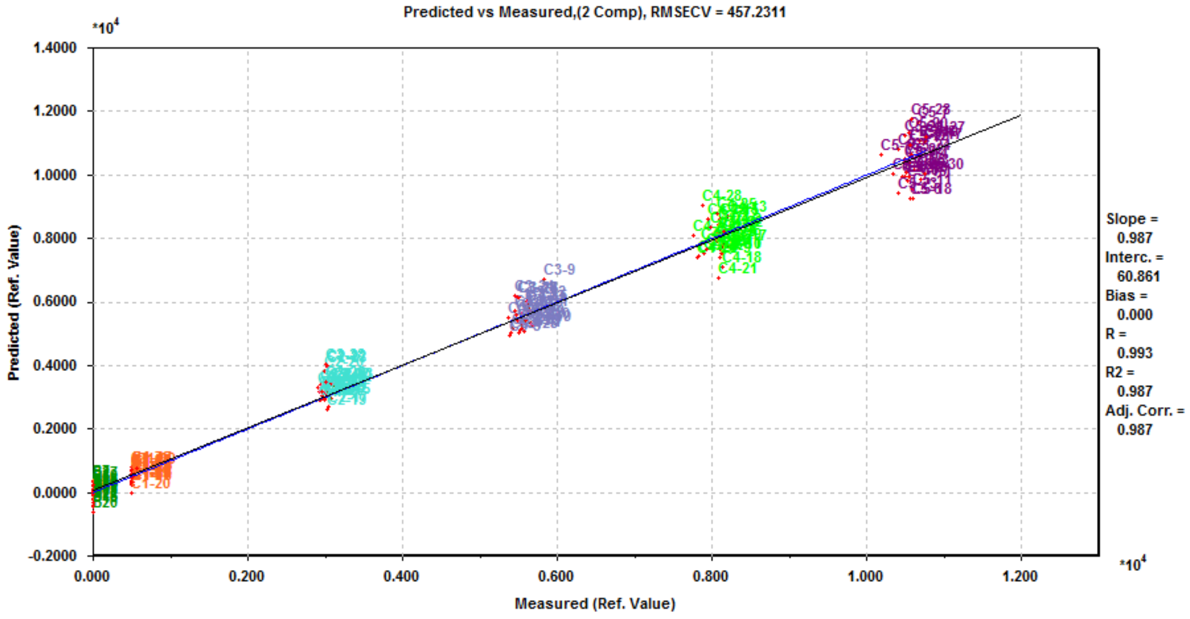


Figure 5.20: Predicted verses measured plot

# 6 Conclusion and Future work

In the rapidly changing and complex field of food authentication, it is becoming increasingly important that rapid and non-destructive techniques are employed to ascertain authenticity.

The results of this study illustrate the ability of NIR spectroscopy together with SIMCA model to discriminate pure wheat gluten samples from adulterated wheat gluten samples. The study indicates 100 % classification rate at adulteration level of as low as 3000 ppm. The SIMCA model indicated overlap at spiked level of 500 ppm. Overlap at 500 ppm was also confirmed by PLSR model. The method is validated at laboratory levels. Further study on these models will include the transfer of these lab-based method into an industrial setting where the incoming batches of materials would be screened. However the model needs to be updated by adding more new samples to account for raw material variation, seasonal quality fluctuation and production variation amongst others.

In the current study only one type of NPN compound has been studied. To develop non target NIR methods to be useful in routine authentication of wheat gluten samples, samples spiked with different NPN compounds needs to be included in the model.

Routine testing of shipment of wheat gluten meal samples with NIRS could have economic benefits for the feed sector.

# References

1. Everstine, K. et al. (2013) Economically motivated adulteration (EMA) of food: common characteristics of EMA incidents. Journal of Food Protection 76 (4), 723-735.

2. Phromkunthong, W. et al. (2013) Toxicity of melamine, an adulterant in fish feeds: experimental assessment of its effects on tilapia. Journal of fish diseases 36 (6), 555-568.

3. Chain, E.P.o.C.i.t.F. et al. (2010) Scientific Opinion on Melamine in Food and Feed. EFSA Journal 8 (4).

4. Haughey, S.A. et al. (2015) The use of handheld near-infrared reflectance spectroscopy (NIRS) for the proximate analysis of poultry feed and to detect melamine adulteration of soya bean meal. Analytical Methods 7 (1), 181-186.

5. Sun, F. et al. (2010) Analytical methods and recent developments in the detection of melamine. TrAC Trends in Analytical Chemistry 29 (11), 1239-1249.

6. Rovina, K. and Siddiquee, S. (2015) A review of recent advances in melamine detection techniques. Journal of Food Composition and Analysis 43, 25-38.

7. Lohumi, S. et al. (2015) A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. Trends in Food Science & Technology 46 (1), 85-98.

8. Cozzolino, D. et al. (2009) Usefulness of near infrared reflectance (NIR) spectroscopy and chemometrics to discriminate between fishmeal, meat meal and soya meal samples. Ciencia e investigación agraria 36 (2), 209-214.

9. Cozzolino, D. et al. (2005) Usefulness of near-infrared reflectance (NIR) spectroscopy and chemometrics to discriminate fishmeal batches made with different fish species. Journal of agricultural and food chemistry 53 (11), 4459-4463.

10. Haughey, S.A. et al. (2013) The application of near-infrared reflectance spectroscopy (NIRS) to detect melamine adulteration of soya bean meal. Food Chemistry 136 (3), 1557-1561.

11. Givens, D.I. et al. (1997) The principles, practices and some future applications of near infrared spectroscopy for predicting the nutritive value of foods for animals and humans. Nutrition research reviews 10 (1), 83-114.

12. Esbensen, K.H. et al. (2002) Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design, Multivariate Data Analysis.

13. Rinnan, Å. et al. (2009) Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 28 (10), 1201-1222.

14. MARTEN, H. and NAES, T., Multivariate Calibration and Classification, Chichester-UK: NIR Publications, 2002.

15. Brereton, R.G. (2003) Chemometrics: data analysis for the laboratory and chemical plant, John Wiley & Sons.

16. Savitzky, A. and Golay, M.J. (1964) Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry 36 (8), 1627-1639.

17. Martens, H. et al., Multivariate linearity transformation for near-infrared reflectance spectrometry, Proceedings of the Nordic symposium on applied statistics, Stokkand Forlag Publishers Stavanger, Norway, 1983, pp. 205-234.

18. Martens, H. and Stark, E. (1991) Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. Journal of pharmaceutical and biomedical analysis 9 (8), 625-635.

19. dos Santos Panero, P. et al. (2013) Application of extended multiplicative signal correction to short-wavelength near infrared spectra of moisture in marzipan. Journal of Data Analysis and Information Processing 1 (03), 30.

20. Næs, T. et al. (2002) A user friendly guide to multivariate calibration and classification, NIR publications.

21. Wold, S. et al. (1987) Principal component analysis. Chemometrics and intelligent laboratory systems 2 (1-3), 37-52.

22. Wold, S. (1976) Pattern recognition by means of disjoint principal components models. Pattern Recognition 8 (3), 127-139.

23. Khanmohammadi, M. et al. (2009) Diagnosis of colon cancer by attenuated total reflectance-fourier transform infrared microspectroscopy and soft independent modeling of class analogy. Medical Oncology 26 (3), 292-297.

24. Rajalahti, T. and Kvalheim, O.M. (2011) Multivariate data analysis in pharmaceutics: A tutorial review. International Journal of Pharmaceutics 417 (1), 280-290.

25. Andersen, C.M. and Bro, R. (2010) Variable selection in regression—a tutorial. Journal of Chemometrics 24 (11-12), 728-737.

26. Rajalahti, T. et al. (2009) Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. Analytical chemistry 81 (7), 2581-2590.

27. Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society. Series B (Methodological), 111-147.

28. Zhang, Y. and Yang, Y. (2015) Cross-validation for selecting a model selection procedure. Journal of Econometrics 187 (1), 95-112.

29. Bradstreet, R.B. (1954) Kjeldahl method for organic nitrogen. Analytical Chemistry 26 (1), 185-187.

30. Grung, B. and Kvalheim, O.M. (1994) Rank determination of spectroscopic profiles by means of cross validation: the effect of replicate measurements on the effective degrees of freedom. Chemometrics and intelligent laboratory systems 22 (1), 115-125.

31. Kruse, S. et al. (1991) Multivariate analysis of proton NMR spectra of serum from rabbits: Monitoring progressive growth of implanted VX-2 carcinoma. Chemometrics and intelligent laboratory systems 11 (2), 191-196.

32. Wold, S. et al. (2001) PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems 58 (2), 109-130.

33. Wold, S. (1978) Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics 20 (4), 397-405.

34. Meza-Márquez, O.G. et al. (2010) Application of mid-infrared spectroscopy with multivariate analysis and soft independent modeling of class analogies (SIMCA) for the detection of adulterants in minced beef. Meat Science 86 (2), 511-519.

35. Candolfi, A. et al. (1999) Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. Journal of pharmaceutical and biomedical analysis 19 (6), 923-935.

36. Wold, S. and Sjöström, M. (1987) Comments on a recent evaluation of the SIMCA method. Journal of Chemometrics 1 (4), 243-245.

37. Rajalahti, T. et al. (2009) Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometrics and Intelligent Laboratory Systems 95 (1), 35-48.