IN-DEPTH INVESTIGATION OF THE EFFECTS OF DIFFERENT PREPROCESSING STRATEGIES ON INFRARED SPECTROSCOPIC DATA

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
In Partial Fullfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Chemistry

by
Elin İlayda DENİZ

December 2021
İZMİR

ACKNOWLEDGEMENTS

I would like to show my deepest gratefulness to people who had been with me during the time I was working on my thesis;

To my supervisor Prof. Dr. Durmuş ÖZDEMİR for his guidance, patience and efforts for my thesis and everything that I have learned from him. It was great honor being able to work with him.

To Prof. Dr. Federico MARINI, for his great expertise and in-depth knowledge given me during my education in La Sapienza University, Rome.

To İnci Holding, AKÇİN and HOSKIN family for their valuable support not only in good times but also in tough times in quarantine in Italy.

To my thesis committee; Prof. Dr. Şerife YALÇIN and Doç. Dr. Levent PELİT for their interest and support.

To my dear friends Merve ARPACIOĞLU and Aylin KENAR for guiding me wisely.

Finaly to my loved ones for always being with me even tho I was stubborn during the challenging times, to my family Şenel and Yavuz DENİZ, to my aunts Sevil BOZTAŞ, Ömür TUNÇ and my uncle Tuncer TUNÇ.

Special thanks to Bahadır SADE, for his great mentorship and for made me realize the value of ‘Live life if there were no second chances’.

ABSTRACT

IN-DEPTH INVESTIGATION OF THE EFFECTS OF DIFFERENT PREPROCESSING STRATEGIES ON INFRARED SPECTROSCOPIC DATA

Whenever collecting experimental data, they may contain several spurious sources of variability which can hinder the extraction of the desired relevant information so that it is rarely the case that they can be processed as such by chemometrics approaches. In recent years, pre-processing techniques has become an integral part of chemometrics modeling with the purpose of providing better endmodels through fundamental knowledge for Near-Infrared (NIR) users. The aim of pre-processing techniques is to improve success of multivariate regression by reducing undesired physical phenomena in the spectra. This thesis describes the theory of present pre-processing techniques and compares the qualitative and quantitative results of their application. Mean centering, scatter-correction methods and spectral derivatives are the instances of those pre-processing techniques used in this thesis. Those techniques and combinations of them have been applied in order to find the best pre-processing strategy. To be able to observe the results and compare the effects of the applied methods, Partial Least Squares and Genetic Inverse Laast Squares are carried out as multivariate calibration methods. When comparing the calibration results of raw data with pre-processing techniques applied data, decreasement standard eror of prediction (SEP) values observed after appliying those techniques, which is good manner. However, better in comparison, t-Test: Paired Two Sample for Means applied. The results demontrates that there is no significant difference within 95% confidence level.

ÖZET

FARKLI VERİ-ÖNİŞLEME STRATEJİLERİNİN KIZILÖTESİ SPEKTROKOPİSİ VERİLERİ ÜZERİNDEKİ ETKİLERİNİN DERİNLEMESİNE İNCELENMESİ

Deneysel veriler toplanırken, istenen ilgili bilgilerin çıkarılmasını engelleyebilecek birkaç sahte değişkenlik kaynağı içerebilirler, bu nedenle kemometrik yaklaşımlarla bu şekilde işlenebilmeleri nadiren olasıdır. Son yıllarda, yakın kızılötesi kullanıcıları için temel bilgiler aracılığıyla daha iyi modeller sağlamak amacıyla, veri ön işleme teknikleri, kemometrik modellemenin ayrılmaz bir parçası haline geldi. Ön işleme tekniklerinin amacı, spektrumdaki istenmeyen fiziksel artıkları azaltarak çok değişkenli regresyonun başarısını arttırmaktır. Bu tez, mevcut ön işleme tekniklerinin teorisini açıklar ve uygulamalarının nitel ve nicel sonuçlarını karşılaştırır. Ortalama merkezleme, saçılım düzeltme yöntemleri ve spektral türevler, bu tezde kullanılan ön işleme tekniklerinin örnekleridir. Bu teknikler ve bunların kombinasyonları, en iyi ön işleme stratejisini bulmak için uygulanmıştır. Sonuçları gözlemleyebilmek ve uygulanan yöntemlerin etkilerini karşılaştırabilmek için kısmi en küçük kareler ve genetik algoritma tabanlı ters en küçük kareler, çok değişkenli kalibrasyon yöntemleri olarak gerçekleştirilmiştir. Ham verilerin kalibrasyon sonuçları, ön işleme teknikleri uygulanan verilerin sonuçları ile karşılaştırıldığında, bu tekniklerin uygulanmasından sonra standart tahminleme hatası (SEP) değerlerinde düşüş gözlemlenmiştir, ki bu sonuçların iyi yönde etkilendiğini göstermektedir. Fakat daha net karşılaştırma yapabilmek amacıyla bağımlı örneklem t-testi uygulanmıştır. Sonuçlar, %95 güven aralığında anlamlı bir fark olmadığını göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

Figure **Page**

LIST OF TABLES

Table **Page**

CHAPTER 1

INTRODUCTION

1.1. Scope of Thesis

During the years, different preprocessing strategies have been proposed, especially for spectroscopic data, where phenomena such as; non-constant baseline, additive effects, multiplicative effects, scattering, and other unwanted phenomena may have a relevant effect on spectra. Despite the large use of spectral preprocessing techniques, literature is still debating on which could be the best pre-processing strategy and how to avoid artifacts, which could damage the information instead of making it more easily accessible.

The purpose of the thesis is to analyze near-infrared (NIR) spectroscopic data with existing preprocessing strategies to verify; which ones have similar effect on the profiles, which improve the signals, which damage it, what could be the best combination of application, in cases where multiple techniques have to be adopted.

1.2. Literature Review

Spectroscopic methods result in multivariate outcomes. These outcomes are reserved and evaluated in order to understand more properties of a product, thus, chemometric models are conducted such as principal component regression (PCA) or partial least square regression (PLSR). Yet, interfering phenomena is often inevitable throughout the measurement, leading to ultimate changes in baseline, position and amplitude of the peaks of spectrograms. In order to avoid that, there is often need for pre-treatment stage for the calibration of chemometric models (Roger et al., 2020).

Ultimately, the pre-processing step aims at improving the subsequent multivariate regression, classification model or exploratory analysis.

Several pre-treatment methods for near-infrared spectral data can be found in literature since they go hand-in-hand (Rinnan et al., 2009). Rinnan et al. states that: "The most widely used pre-processing techniques can be divided into two categories: as scatter-correction methods and spectral derivatives." While scatter corrections include Multiplicative Scatter Correction (MSC), Inverse MSC, Extended MSC (EMSC), Extended Inverse MSC, de-trending, Standard Normal Variate (SNV) and normalization, spectral derivatives include Norris-Williams (NW) derivatives and Savitzky-Golay (SG) polynomial derivative filters.

The underlying trick behind the both types are the decrease of the detrimental effect on the signal-to-noise ratio, which lead to improved subsequent exploratory analysis, bi-linear calibration model, and classification model (Rinnan et al., 2009). Rinnan et al. illustrates the application of MSC, EMSC, SNV, and SG for pectin data. and concludes that it could be more beneficial to use normalization for short-wave NIR-transmission spectra and to use MSC (with first-order reference correction) or SNV for most other case according to their study.

Moreover, the removal of a baseline, the reduction of multiplicative effects, high frequency noise reduction, the correction of harmful spaces, highlighting of spectral details are the instances of them. To begin with, the removal of a baseline is obtained by calculating suitable approximations and then subtracting it from the measured data. Barnes et al. demonstrates that by using a polynomial filter (Barnes et al., 1994) whereas that is achieved by using a low-frequency filter (Eilers, 2004). Another example is the reduction of multiplicative effects. Logaritmic transformation and normalization can be applied for that purpose. Rabatel et al. and Isaksson et al. and Barnes et al. points out those instances (Rabatel et al., 2020), (Isaksson & Næs, 1988), (Barnes et al., 1994). Additionally, high frequency noise reduction is also another example of pre-treatment of calibration of chemometrics. Savitzky et al. displays the reduction of high frequency noise by application of low-pass filtering. Furthermore, Savitzky et al. also indicates how to correct harmful spaces by an orthogonal projection method whilst Andrew et al. İllustrates it by applying Transfer by Orthogonal Projection. Asides from the examples mentioned above, highlighting of spectral details can be found in literature. Roger et al. state that it can be done SG algorithm by excluding the integration of noise magnification.

It is stated in the literature that it is mostly difficult to have a proper choice of pre-treatment method since there are variety of choices. Therefore, it is possible to find articles combining different pre-treatment methods (Roger et al., 2020). It is also mentioned that the order or application of these pre-treatment methods might have an impact on the success of the process (Engel et al., 2013). Thus, trial and error procedure is the ultimate way when it comes to combination of different pre-treatments, which is done sequentially. In order to rationalize this trial and error loop, there are studies presenting various methodologies such as ensemble learning/ ensembling method, sequential preprocessing through orthogonalization and etc (Dietterich, 2000) (Ni et al., 2009).

Rinnan et al. mentions that in principle, any sequence of pre-processing is possible for combinations of pre-processing methods in literature. According to the study, there are certain points to be considered and not every combination result in success.
Kawamura et al. points out that the use of Genetic Algorithm-Based Partial Least Squares Regression accompanied with using the Savitzky-Golay smoothing filter and a standard normal variate transform (SNV) was employed to reduce the particle size effect for the assessment of Soil Phosphorus Content (Kawamura et al., 2019). Laxalde et al demonstrates the development of chemometric predictive method of near infrared (NIR) spectra for the quantitative determination of saturates, aromatics, resins and asphaltens in heavy petroleum products (Laxalde et al., 2011). Genetic algorithm for the optimisation or co-optimisation of pre-processing and variable selection is evaluated.

CHAPTER 2

INSTRUMENTATION

In 1800, Sir Frederick William Herschel (1738–1822) found Infrared region (IR) of electromagnetic spectrum. The infrared region is a wide spectral region which comprise radiation with wavelength from 0.78 μm to 1000 μm and wavenumber from 12,800 cm^{-1} to 10 cm^{-1}. According to instrumentation and usage area, IR spectrum is subdivided into three regions which are Near Infrared (NIR), Middle Infrared (MIR), and Far Infrared (FIR) radiation as shown in Table 2.1. in terms of wavelength (λ) and wavenumber (v) (Skoog et al., 2017).

Table 2.1. Spectral regions of infrared radiation.

Region	Wavelength (λ), μm	Wavenumber (v), cm^{-1}
Near	0.78 to 2.5	12800 to 4000
Middle	2.5 to 50	4000 to 200
Far	50 to 1000	200 to 10
Most used	2.5 to 15	4000 to 670

Infrared spectroscopy is absorption or transmission of infrared radiation by sample by analyzing the vibrations within atoms of the molecule of interest. Each molecular structure has characteristic vibration. Thus, polyatomic molecules give rise to unique spectrum and the spectrum of vibrational energies provides qualitative analysis.

Before Fourier Transform was introduced, only qualitative analysis, structural identification, was possible by mid-IR region spectrometers which were dispersive type instruments. IR spectrometers were based on grating monochromators or prism. However, in 1980s, with the invention of interferometer by Albert Abraham Michelson,

quantitative analysis also became possible by using Fourier Transform Infrared (FT-IR) spectrometry. Simple block diagram of FT-NIR spectrometer is shown in Figure 2.1.

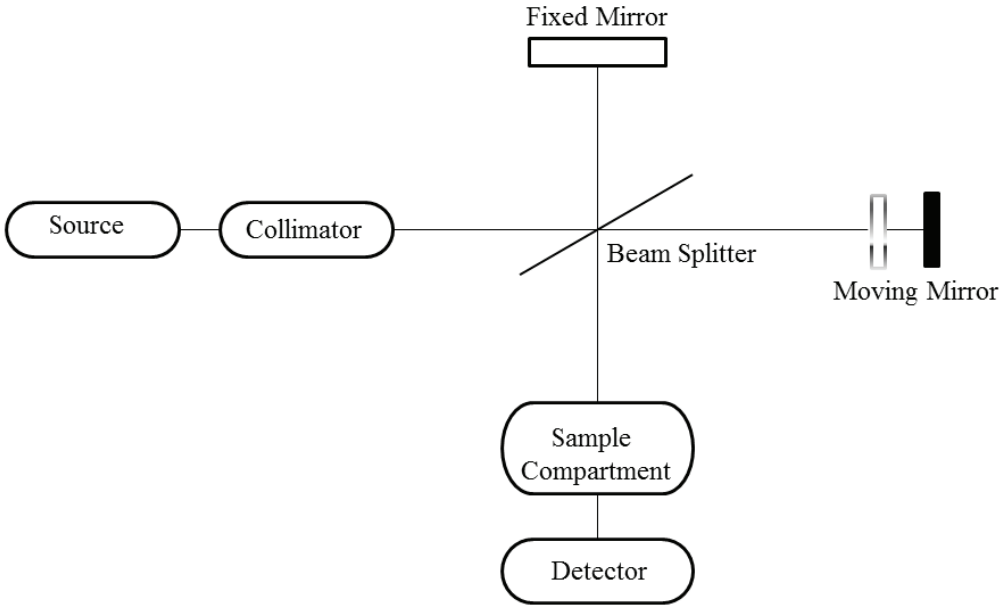


Figure 2.1. Simple block diagram of Fourier Transform Near Infrared (FT-NIR) spectrometer.

Infrared spectrum which is collected from middle infrared region of fourier transform type infrared spectrometer is divided into two parts, functional groups region (3600-1200 cm^{-1}) and fingerprint region (1200-600 cm^{-1}). Furthermore, the spectral information obtained from vibrational and rotational stretching modes provide identification of unknown material while absorbance values are linearly correlated with concentration and allows to determine number of components in a mixture.

The benefits of using FT-IR spectrometer over dispersive type infrared spectrometers and other techniques can be listed as:

- It has auto calibration of wavelength,
- It has multiplex advantage which means many scans can be done and averaged in a shorter time,
- It uses scan averaging to enhance signal to noise (S/N) ratio,
- It has high resolution,

- It is a non-destructive technique,
- It has good precision,
- It is ease of use and data can be reproducible.

Therefore, FT-IR spectroscopy allows both qualitative and quantitative analysis and as result, processing FT-IR spectroscopy with chemometrics methods provide successful classification and calibration.

NIR spectroscopy divided into two categories as transmission spectroscopy and diffuse reflectance spectroscopy. While transmission spectroscopy is applicable for liquids, diffuse reflectance spectroscopy is applicable for solids. Transmittance (T) and diffuse reflection (R) are given below in Equation 2.1. and 2.2. respectively (Pasquini, 2003), (Stuart, 2004).

$$T = \frac{I}{I_0} \tag{2.1}$$

Where intensity of incoming light is I_0 and intensity of transmitted light is I.

$$R = \frac{I}{I_R} \tag{2.2}$$

Where I_R is the light intensity which is reflected from surface and I is the light intensity reflected from solid sample.

Since the diffuse reflection of different wavelengths of light is being measured in diffuse reflectance spectroscopy in order to procure information from the surface of solids, particle size is being critical for quantitative chemometrics models (Dai et al., 2018). Another critical point of view for diffuse reflectance spectroscopy is multi-collinearity problem. Multiplicative combinations of particle size effect and multi-collinearity are one of the important problems which inhibit the interpretation of the spectra of NIR diffuse reflectance (Barnes et al., 1989).

On the other hand, using NIR technology to predict physical and chemical information is being beneficial. So that NIR technology is useful for quantification of adulterants (Dai et al., 2018).

CHAPTER 3

PRE-PROCESSING TECHNIQUES

Chemometrics approaches are mathematical methods used to provide relevant chemical information by analyzing chemical data. So that, collecting optimal data and pre-processing of spectral data are very important process to construct better chemometric model. Pre-processing techniques can be used to improve both subsequent exploratory analysis, classification model or bi-linear calibration model to make data obey Beer's Lambert law. As being the empirical equation for NIR, Beer's Lambert law propose linear relation between the absorbance and concentration of the components of interest, as given in Equation (3.1) (Rinnan et al., 2009).

$$\mathbf{A}_\lambda = \boldsymbol{\varepsilon}_\lambda . \mathbf{b} . \mathbf{c} \qquad (3.1)$$

Here, $\mathbf{A}_\lambda$ and $\boldsymbol{\varepsilon}_\lambda$ represent wavelength dependent absorbance and molar absorbtivity respectively, while **c** is the concentration and **b** is the path length of light through sample. Comparable size of wavelengths in NIR electromagnetic radiation and particle sizes of samples can cause light scattering. So that undesired effects in spectral data can be occur. It is important to eliminate undesired effects before constracting a calibration model by applying suitable pre-processing. The most common used pre-processing techniques are Mean Centering, Baseline Correction, Scatter Corrections and Spectral Derivatives. While as scatter correction methods Multivariate Scatter Correction, Extended Multivariate Scatter Correction, Standard Normal Variate, and as spectral derivatives Savitzky-Golay Polynomial Derivatives are performed in this thesis (Rinnan et al., 2009).

3.1. Mean Centering

Mean centering is required for many multivariate calibration models either to reduce multicollinearity or improve interpretation of the resulting regression equations. Mean centering of the columns done by subtracting the mean of each column (or variable) as given in Equation 3.2 (Brereton et al., 2003).

$$^{cen}\mathbf{x}_{ij} = \mathbf{x}_{ij} - \bar{\mathbf{x}}_j \tag{3.2}$$

3.2. Baseline Correction

The strategy of baseline correction is fitting a line to account for baseline followed by the subtraction. Construction of a least squares model by regressing **X** with absorbance values for each spectra is as shown in Equation 3.3 (Akkoç, 2018)

$$\boldsymbol{a} = \mathbf{X} \,.\, \boldsymbol{\beta} + \mathbf{e} \tag{3.3}$$

Where a is the single vector containing absorbance values for a single spectrum.

$$\boldsymbol{b} = \mathbf{X} \,.\, \widehat{\boldsymbol{\beta}} \tag{3.4}$$

b refers to the baseline to be subtracted from the spectrum in Equation 3.4.

3.3. Scatter Corrections

The unique chemical characteristic of sample allows different absorption at different wavelengths of the incident light, and it results with spectra related to analyte of interest. However, scattering effects which are additive effects, occurs as baseline displacement on spectra that caused by path length difference, and multiplicative effects, occurs as change in slope of spectrum caused by differences in particle size, cause

variations in NIR spectra. So that, the NIR spectra contains mixture of absorbance and scatter (Rinnan et al., 2009).

Scatter Correction techniques are designed to transform the spectra related to scatter artifacts into the spectra related to analyst of interest, in a way that reducing variability due to scatter. The most common used scatter correction methods are Multiplicative Scatter Correction and Standard Normal Variate.

3.3.1. Multiplicative Scatter Correction (MSC)

Multiplicative Scatter Correction (MSC) was firstly introduced by Martens et al. in 1983, then detailed by Geladi et al. in 1985 and became a commonly used scatter removal technique for NIR (Martens et al., 1983), (Geladi et al., 1985). The aim of MSC is correcting both additive and multiplicative effects on spectral data corresponds to the reference spectrum which is generally average spectrum of the calibration set (Mishra et al., 2020). The correction is done by changing the scale and the offset of the spectra and results with relatively consistent baseline. Estimation of the slope and offset terms by utilizing least squares regression is given in equation 3.5 and correction of original spectrum is given in equation 3.6.

$$\mathbf{A}(\boldsymbol{\lambda}) = \boldsymbol{a} + \bar{\mathbf{x}}\,(\boldsymbol{\lambda}).\boldsymbol{b} + \boldsymbol{e}(\boldsymbol{\lambda}) \tag{3.5}$$

Here $\mathbf{A}(\boldsymbol{\nu})$ is the measured absorbance, $\bar{\mathbf{x}}$ is mean of all spectra, $\boldsymbol{\lambda}$ is wavelength, a and b are coefficients.

$$\mathbf{A}_{\mathbf{corr}}(\boldsymbol{\lambda}) = (\mathbf{A}(\boldsymbol{\lambda}) - \boldsymbol{a})/\boldsymbol{b} \tag{3.6}$$

3.3.2. Extended Multiplicative Scatter Correction

In 1989, Extended Multiplicative Scatter correction (EMSC) developed as elaborated version of the MSC by Stark and Martens (Afseth & Kohler, 2012). EMSC is

used to model data with using priori information from the spectra of interest by removing scattering effects caused by physical phenomena. Besides, it overcomes with baseline displacement by fitting of baseline on the wavelength axis. Since EMSC is a higher order expansion of the MSC, it comprises second order polynomial fitting to the reference spectrum as shown in equation 3.7.

$$\mathbf{A}(\boldsymbol{\lambda}) = a + \bar{\mathbf{x}}\,(\boldsymbol{\lambda}).b + d_1.\boldsymbol{\lambda} + d_2.\boldsymbol{\lambda}^{\mathbf{2}} + \cdots + d_n.\boldsymbol{\lambda}^{\mathbf{n}} \qquad 3.7$$

Here d_1 and d_2 are coefficients found from regression. Correction of original spectrum is as given below in equation 3.8.

$$\mathbf{A_{corr}}(\boldsymbol{\lambda}) = (\mathbf{A}(\boldsymbol{\lambda}) - a - d_1.\boldsymbol{\lambda} + d_2.\boldsymbol{\lambda}^{\mathbf{2}} + \cdots + d_n.\boldsymbol{\lambda}^{\mathbf{n}})/b \qquad (3.8)$$

3.3.3. Standard Normal Variate (SNV)

Standard Normal Variate (SNV) is another most widely applied scatter correction technique which is used for normalizing NIR data to reduce both additive and multiplicative effects. The difference with MSC is that SNV does not require reference spectrum. The correction is done by centering each spectrum and dividing by its standard deviation as shown in Equation 3.9 (Rinnan et al., 2009).

$$\mathbf{x}_{\mathbf{i,j}}^{\mathbf{SNV}} = \frac{\mathbf{x}_{\mathbf{i,j}} - \overline{x_i}}{s(x_i)} \qquad (3.9)$$

Where $\bar{x}_i$ is the mean of spectrum i and $s(x_i)$ is the standard deviation of spectrum.

Since whole spectrum divided by standard deviation, changing in just a part of spectrum can impact entire of spectrum. It causes mismatching between coefficients and physical information in the spectra. So that it effects the robustness off model.

3.4. Spectral Derivatives

The principle of spectral derivatives is taking derivative of absorbance with respect to wavelength. Both first and higher order derivatives used to remove multiplicative and additive effects in the spectra by smoothing. While the first derivation of spectral data removes only the baseline, the second derivation have capability to remove both baseline and linear trend. Zero, first and second order derivatives of spectra can be present like in Equations 3.10, 3.11 and 3.12, respectively (Brereton, 2003).

$$f(\boldsymbol{\lambda}) = \mathbf{A} \tag{3.10}$$

$$f'(\boldsymbol{\lambda}) = \frac{d\mathbf{A}}{d\boldsymbol{\lambda}} \tag{3.11}$$

$$f''(\boldsymbol{\lambda}) = \frac{d^2\mathbf{A}}{d\boldsymbol{\lambda}^2} \tag{3.12}$$

Where, the spectrum is expressed as absorbance, A, as a function of wavelength.

3.4.1. Savitzky-Golay (SG) Polynomial Derivatives

Savitzky and Golay proposed a very efficient method to calculate the derivative by fitting polynomial in a window size of 2n+1 data, at a particular wavelength. This operation is applied to all points in the spectra sequentially. The least square solution for fitting polynomial is given in Equation 3.13 (Brereton, 2003).

$$\mathbf{A}_{\boldsymbol{\lambda}} = a_0 + a_1\boldsymbol{\lambda} + \mathbf{K} + a_1\boldsymbol{\lambda}^{\boldsymbol{l}} \tag{3.13}$$

Here, l is the degree of the fitted polynomial. When the parameters for the l order polynomial are calculated, it allows to take the derivative of any order of the function.

CHAPTER 4

MULTIVARIATE CALIBRATION METHODS

Calibration method can be explained as relation of predictor variables to response variables obtained by regression techniques which build up with mathematical model, in the field of chemistry (Brereton, 2003). The main goal of regression techniques is yielding better prediction by the builded mathematical model. The efficiency of the calibration method can be effected from the type of mathematical model (i.e. linear, quadratic, exponential) and features of collected data which are number of variables and samples, correlation of variables, unwanted phenomenones like noise (Akkoç, 2018). However, the characteristics of Near-Infrared (NIR) spectroscopic data results wide and overlapping peaks (Skoog et al., 2017).

Multivariate calibration relies on multivariate variables and taken into consideration as an extension of Beer-lambert's Law for determination of concentration from absorbance data. Multivariate calibration methods take adventages of addressing not only to multiple variables but also multiple responses for prediction in the case of the compond of interest take shape with different wavelength of maximum absorbance each refers different properties. However, multivariate calibration methods require some important limitations to be able to obtain consistent model. These limitations can be listed as:

- Observations' quantity should exceed the variables.
- Variables' quantity should be equal to at least the compounds (Meşe, 2016).

4.1. Classical Least Squares (CLS)

As a multivariate extension of Beer Lambert's Law, Classical Least Squares (CLS) is one of the multivariate calibration method which aims to predict concentration

while taking absorbance value as function of it, as shown in Equation (4.1) (Şentürk, 2020).

$$\mathbf{A} = \mathbf{C} \,.\, \mathbf{K} + \mathbf{E} \tag{4.1}$$

In Equation (4.1), **A** is *nxl* matrix of absorbance values which contains the predictor variables for concentrations, **C** is the *nxm* matrix of concentration values that are response variables of calibration set where. Here, *n* is the number of samples and l is the number of wavelengths in related spectrum, m is the number of components which refers number of response variables related to predictor variables. Solving equation for unknown K which is *mxl* matrix of absorptivity coefficients, requires construction of spectra by using concentrations. Therefore as a limitation of CLS, concentration of each species, which has a significant impact on the spectrum, must be known to be able to have successful model. The least squares solution is given in Equation (4.2).

$$\widehat{\mathbf{K}} = (\mathbf{C}'.\mathbf{C})^{-\mathbf{1}}.\mathbf{C}'.\mathbf{A} \tag{4.2}$$

After **K** matrix is obtained, prediction for concentration of components in an unknown sample is obtained by

$$\hat{\mathbf{c}} = (\widehat{\mathbf{K}}.\widehat{\mathbf{K}}')^{-\mathbf{1}}.\widehat{\mathbf{K}}.\mathbf{a} \tag{4.3}$$

Where **a** is *lx*1 vector absorbance values obtained from the unknown species' spectrum and **c** is *mx*1 vector for concentrations of component in unknown species.

CLS method gives successful results when concentration of each species is known. Errors (**E**), that are caused by instrument, consist of residuals of absorbance which are not fitted to equation of builded model. The precision and accuracy of instruments are improved over the last decades. However; preparation of solutions with different concentrations results with errors which caused by human involving process and make precision remain limited. Therefore, there are many reasons to suppose that the source of error is from the concentration, as in ILS approach reported in following section.

4.2. Inverse Least Squares (ILS)

Inverse Least Squares (ILS) is a multivariate extension of inverse Beer Lambert's Law which takes concentration as a function of absorbance, as in Equation (4.4) (Şentürk, 2020).

$$\mathbf{C} = \mathbf{A} \,.\, \mathbf{P} + \mathbf{E} \tag{4.4}$$

While **C** and **A** refers to same matrices as described in the CLS, **P** is *lxm* matrix which represents regression coefficients that relate absorbance values with component's concentration. Errors (**E**) consist of residuals of concentration. When **E** is *nxm* matrix which assumed to be independent, identical analysis for each analyte can be done by modeling a single component at a time as shown in Equation (4.5)

$$\mathbf{c} = \mathbf{A} \,.\, \mathbf{p} + \mathbf{e} \tag{4.5}$$

The calculation of **P** with using pseudo-inverse is given in Equation (4.6).

$$\mathbf{P} = (\mathbf{A}'.\mathbf{A})^{-\mathbf{1}}.\mathbf{A}^{\mathbf{l}} \times \mathbf{c} \tag{4.6}$$

Equation (4.7) shows that even without information of other species, it is possible to model one component in a sample at a time.

$$\hat{\mathbf{p}} = (\mathbf{A}'.\mathbf{A})^{-\mathbf{1}}.\mathbf{A}^{\mathbf{l}} \times \mathbf{c} \tag{4.7}$$

While vector $\mathbf{p}_{mx1}$ refers to regression coefficient, vector $\mathbf{c}_{nx1}$ refers to concentrations of the modelled component. After obtaining **p**, unknown sample is predicted as shown in Equation (4.8).

$$\hat{c} = \mathbf{a}'.\hat{\mathbf{p}} \tag{4.8}$$

Even though ILS approach takes advantages of modelling single component at a time, it becomes evident that pseudo-inverse calculation causing multicollinearity problem. While higly correlated data provide best fit to calibration set, overfitting of the model on training set can be observed. To overcome with the multicollinearity problem, absorbance values which are irrelevant to concentrations can be leave out by performing

Partial Least Squares aproach which is factor based projection or Genetic algorithms which are feature selection methods.

4.3. Partial Least Squares (PLS)

As one of the projection method, Partial Least Squares (PLS) provides prediction of multicompound and solves multicollinearity problem by explain whole data with a few latent variables. Since PLS projects absorbance and concentration data to a new dimensional space in such way that maximizing covariance between them, it is advantagable as ensuring information of the projected variables which explain responses. Moreover, PLS assumes that errors from both absorbance and concentrations (Geladi et al., 1986). PLS decomposes both **A** matrix and **c** vector, as shown in Equation (4.9) and (4.10), respectively.

$$\mathbf{A} = \mathbf{T} . \mathbf{B} + \mathbf{E} \tag{4.9}$$

$$\mathbf{c} = \mathbf{U} . \mathbf{r} + \mathbf{e} \tag{4.10}$$

Here, *h* represents the Latent Variables LVs (number of components: PC). The term **A** is *mxn* matrix refers to the responses where **T**, *m*x*h* matrix, is the **A** scores, **B** is *h*x*n* matrix which are loadings and **E** is *mxn* matrix which is the residuals of **A**. The term **c** is is *m*x1 vector and refers to the property of interest where **U**, *m*x*h* matrix, is the scores and **r**, *hxl* matrix, is the loadings of **c**, which obtained by solving given equations iteratively, and **e** is *m*x1 vector the residuals of **c**. Prediction by multiplication of **A** scores with **c** loadings shown in equation (4.11).

$$\hat{\boldsymbol{c}} = \mathbf{T} . \hat{\mathbf{r}} \tag{4.11}$$

Selection of optimum number of LVs is one of the significant step to build successful model. For PLS, a plot of number of Latent Variables (LVs) vs. corresponding Predicted Residual Error Sum of Squares (PRESS) is used to select the optimal LVs. The calculation of PRESS given below in Equation (4.12).

$$PRESS = \sum_{i=1}^{n} (\hat{c}_i - c_i)^2 \quad (4.12)$$

Here, $\hat{c}$ is the predicted concentration of component. The number of optimal LVs can be choosen at the point where PRESS starts to increase or at the point where PRESS values stop decreasing. Selection of extra LVs is likely to make the model prone to overfitting.

4.4. Genetic Inverse Least Squares (GILS)

In 1960, Genetic Algorithm (GA) was introduced as global search and Viaoptimization method by John Holland with the inspiration of Charles Darwin's theory of natural evolution and selection. GA allows not only optimization of the regression coefficients, which minimize error of prediction, but also finding best combination of variables that are well correlated with response variables. Since these combinations increase exponentially with number of variables, more than 3000 in NIR spectra, it is not feasible to test each combination. So that GA created to solve this problem to be able to make fitness (inverse of standard error of cross validation) computationally quickly.

Genetic Inverse Least Squares (GILS) combines GA for iterative variable selection and ILS for building calibration model (Özdemir & Öztürk, 2004).

GA come up with 5 main steps which are listed below:

1. Initialization of gene pool
2. Evaluation of the genes in the population
3. Selection of parent genes for breed
4. Cross-over of genes
5. Replacement of parent genes by off-springs

4.4.1. Initialization of Gene Pool

GA starts with initialization of gene pool, where a gene refers to a combination of variables which should not exceed number of samples. Representation of a gene which expressed as G, is given in Equation 4.12.

$$\mathrm{G}: [\mathrm{A}_{812}, \mathrm{A}_{2703}, \mathrm{A}_{758}, \mathrm{A}_{922}, \mathrm{A}_{1881}, \mathrm{A}_{3439}] \tag{4.12}$$

Here, A is the absorbance at i^{th} wavelength. Leave one-out cross validation is used for calculation of the fitness and R^2 of the gene. Genes with the R^2 value of higher than threshold (e.g., 0.5) are accepted to the population. This process continues till the required number of genes are selected to the population.

4.4.2. Evaluation of The Genes in The Population

The success of the gene which is selected to the initial gene pool is evaluated by using fitness function, given in Equation 4.13, which is inverse of the Standard Error of Cross Validation (SECV).

$$\mathrm{FITNESS} = \frac{1}{\mathrm{SECV}} \tag{4.13}$$

The success of each gene is indicated by SECV which is given in Equation 4.14.

$$\mathrm{SECV} = \sqrt{\frac{\sum_{i=1}^{m}(c_i - \hat{c}_i)^2}{m-2}} \tag{4.14}$$

Here, m is the number of samples, c_i is the reference concentration while $\hat{c}_i$ is the predicted concentration. When the calculation of the fitness values of each gene is done, the genes are sorted by decreasing order of their fitness.

4.4.3. Selection of Parent Genes for Breed

Breeding requires exchange of information between genes, thus, selection of gene pairs. Since there are several selection methods as tournament and top-down, roulette wheel selection was used. The visualization of roulette wheel selection is as given in Figure 4.1. Each area on the wheel matches with the particular gene. Since the largest one on the roulette wheel represents the gene with the highest fitness value, probability of the being selected of the gene is also higher when the wheel is spun.

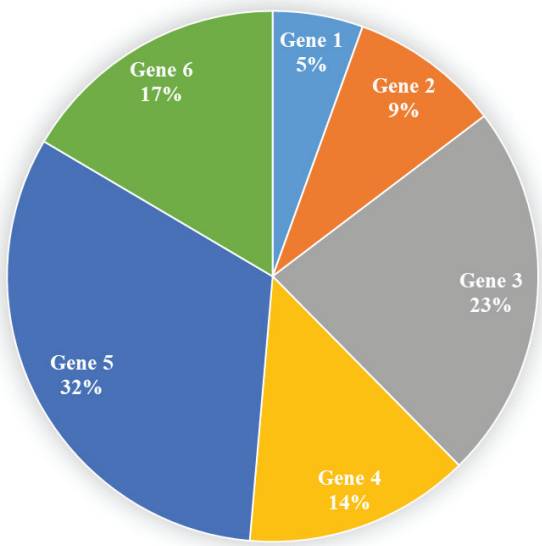


Figure 4.1. Visualization of Roulette Wheel.

4.4.4. Cross-Over of Genes

Off-Spring genes are created by cutting parent genes and combining first half of the gene with the last half of the other gene. The illustration of cross-over is given below. While parent genes are represented as G1 and G2, off-springs are NEW G1 and NEW G2. The | symbol shows where the genes are cut off.

Parent Genes:

G1: $[A_{812}, A_{2703}, A_{758} \mid A_{922}, A_{1881}, A_{3439}]$

G2: $[A_{215}, A_{2703} \mid A_{623}, A_{489}]$

Off-Spring Genes:

NEW G1: $[A_{812}, A_{2703}, A_{758}, A_{623}, A_{489}]$

NEW G2: $[A_{215}, A_{2703}, A_{922}, A_{1881}, A_{3439}]$

Fitness values of each gene are also determined with concluding in cross-over of all selected gene pairs.

4.4.5. Replacement of Parent Genes by Off-Springs

Replacing parent genes with off-springs done in each iteration without considering their fitness values. Then, fitness of each off-springs is compared with the fittest gene of the previous iteration. If the off-spring in the new generation is fitter, then, it is stored. This step is repeated till reaching the defined number of iterations. The procedure starting from initialization and to the end of last iteration is called a run. With the purpose of constructing a model, the fittest gene is saved at the end of each run. By using the constructed model obtained from GILS, concentrations of independent validation data set can be predicted. The standard error of prediction (SEP) is calculated to examine the success of the model as given below in equation 4.15.

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{m}(c_i - \hat{c}_i)^2}{m}} \tag{4.15}$$

Where, m is number of samples in the validation data set.

CHAPTER 5

DATA PROCESSING

5.1. Data Analysis

Determination of quality parameters such as protein, moisture, and starch content of agricultural food products by wet chemistry analyses takes long time. Near infrared spectroscopy (NIR) coupled with multivariate calibration offers a fast and nondestructive alternative to obtain reliable results. However, due to the complexity of the spectra obtained from NIR, some wavelength selection is generally required to improve the predictive ability of multivariate calibration methods. In this study, Cargill corn data sets are investigated with the aim of establishing successful calibration models using NIR spectra of corn samples (Tan & Brown, 2003). The data set was downloaded from http://software.eigenvector.com/Data/ Corn/index.html and contained 80 NIR spectra of corn of which wet chemical analysis of protein, moisture, oil, and starch content were done with reference methods. Outputs for those contents are from 3 different instrument which are m5, mp5 and mp6. The spectra were recorded in diffuse reflectance mode as log (1/R) and the wavelength range was between 1100 and 2498 nm, measured at 2 nm intervals. Only the spectra for protein content of instrument m5 were selected as data set for this thesis. Histogram chart of reference protein (w/w%) is as given below in Figure 5.1. Protein values vary between 7.654 and 9.711 as (w/w%).

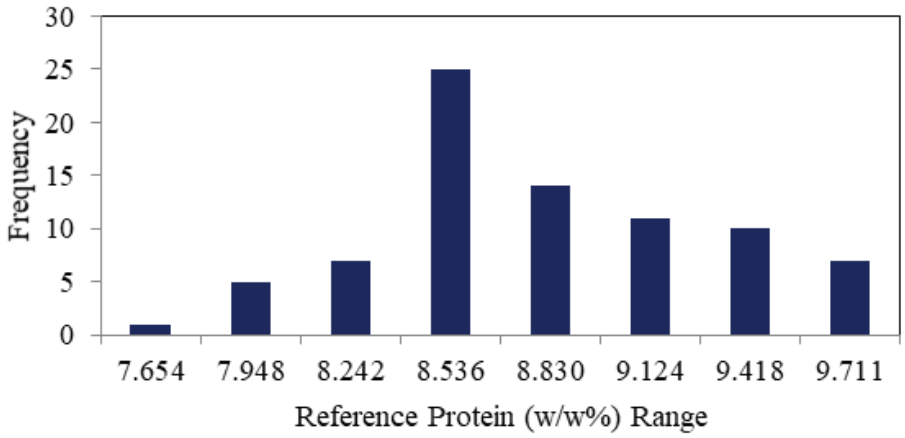


Figure 5.1. Histogram chart of reference protein (w/w%)

60 of the NIR spectra were included in calibration set and 20 of them were included in validation set. Calibration Toolbox from Oba Kemometri is used for preprocessing, evaluation, extraction, and modeling of data. Preprocessing Toolbox from Oba Kemometri is used for preprocessing the data, examining, and removing the problematic data before modeling, separating the whole data set as a calibration and validation set. (URL-1)

In the scope of the thesis, different preprocessing techniques and their combinations are applied on protein FT-NIR data. The effect of each method was criticized with the consideration of results of evaluated models by performing PLS and GILS multivariate calibration techniques.

5.2. Preprocessing, Selection Approaches and Limitations

To achieve successful model, selecting suitable pre-processing is crucial. The goal of pre-processing techniques is to remove un-wanted phenomenon in the data to enhance the feature sought in the spectra. This can be achieved by using a suitable pretreatment. However, applying the wrong type or applying a too severe preprocessing can cause the loss of valuable information. The selection of proper pre-processing is important to assess prior to model validation. Pre-processing should maintain or decrease the effective model

complexity. Generally, performing several pre-processing steps is not advisable. The purpose of this study is to demonstrate effects of different pre-treatments and different combinations of them on the FT-NIR corn data. The list of pretreatments applied on the corn NIR dataset is given below in Table 5.1.

Table 5.1. List of the Pre-processing Techniques applied on the corn NIR dataset and their abbreviations.

Pre-processing Technique	**Abbreviation**
Mean Centering	MC
Baseline Correction	BC-P
Standard Normal Variate	SNV
Multiplicative Scatter Correction	MSC
Extended Multiplicative Scatter Correction	EMSC
Savitzky Golay	SG-W-P-D

In the table 5.1., MC, S, SNV, MSC and EMSC mean Mean Centering, Standard Normal Variate, Multiplicative Scatter Correction and Extended Multiplicative Scatter Correction respectively. BC-P means Baseline Correction using an P^{th} order polynomial and SG-W-P-D means Savitsky and Golay using a W points wide window, an P^{th} order polynomial and a D^{th} derivative.

In order to observe the qualitative effect of preprocessing techniques on data, spectral visualization of FT-NIR corn data and preprocessing techniques applied data are given as follows. Figure 5.2. show the FT-NIR spectra of corn raw data.

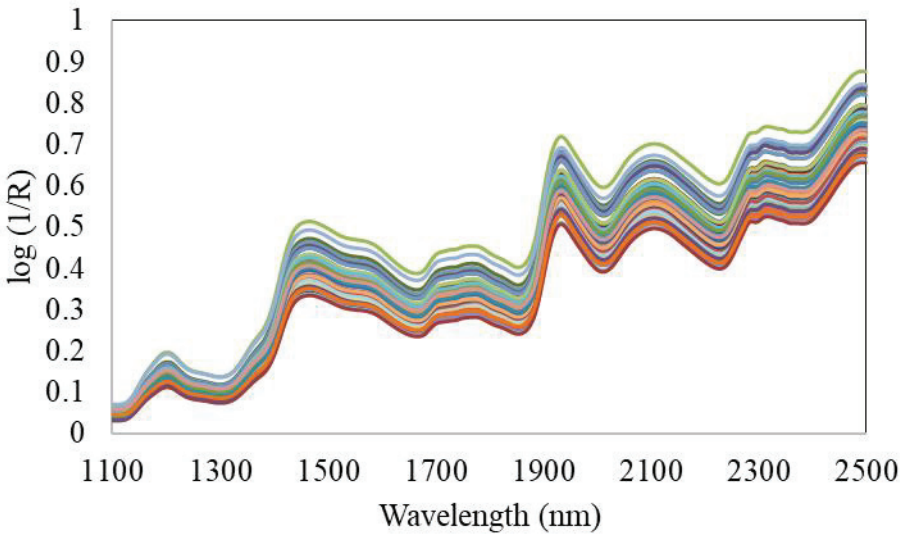


Figure 5.2. FT-NIR spectra of corn raw data.

As a dimensional approach of mean centering is applied to corn data by subtracting mean of the spectrum from each spectrum, so that the new mean of spectrum is as shown in Figure 5.3.

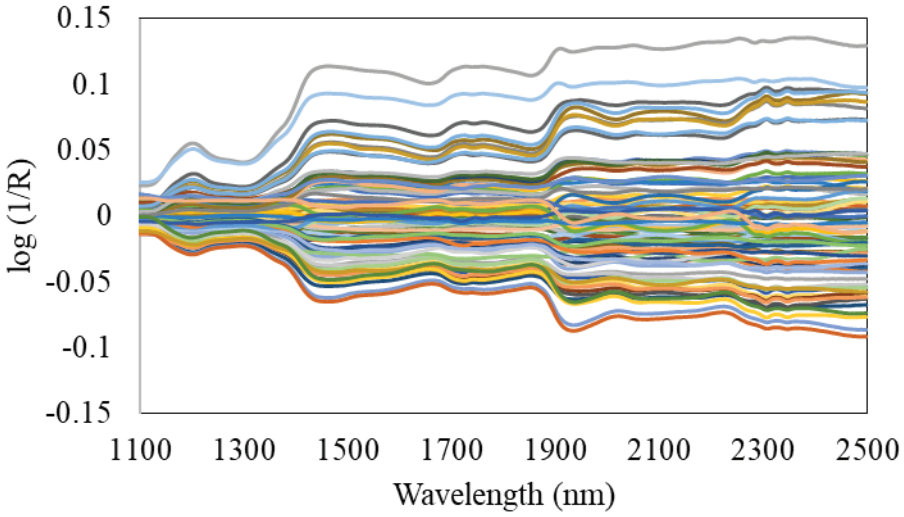


Figure 5.3. FT-NIR spectra of mean centering applied data.

Baseline Correction is done by fitting a line to wavelength with the selected points of 100; 375; 565 and substracted from spectra. Representation of FT-NIR spectra of 1st polynomial order Baseline Correction applied corn data and FT-NIR spectra of 2nd polynomial order Baseline Correction applied corn data are as shown in Figure 5.4 and 5.5, respectively.

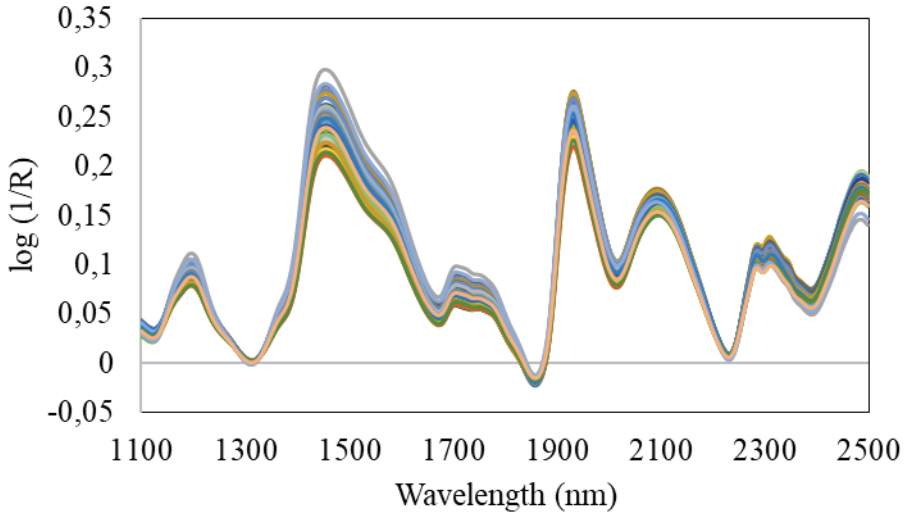


Figure 5.4. FT-NIR spectra of 1st polynomial order Baseline Correction applied corn data.

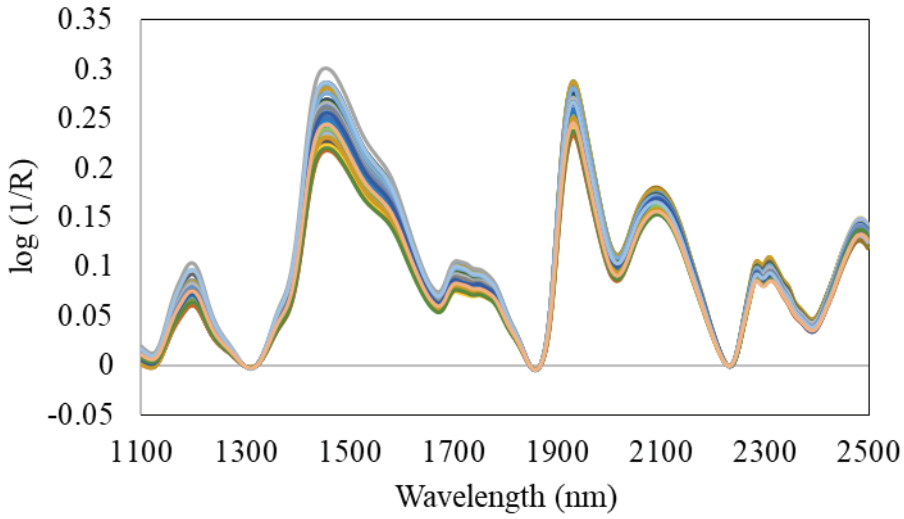


Figure 5.5. FT-NIR spectra of 2nd polynomial order Baseline Correction applied corn data.

Standard Normal Variate as being one of the scatter correction method is performed on corn data set. Figure 5.6. demonstrates the SNV correction for the corn dataset.

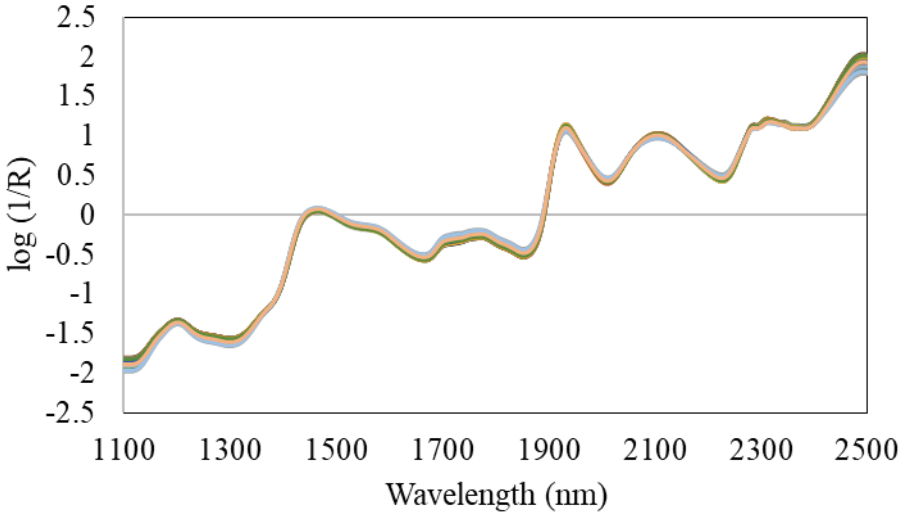


Figure 5.6. FT-NIR spectra of Standard Normal Variate applied corn data.

Another scatter correction method Multiplicative Scatter Correction is used to remove variance caused by scattering, especially in spectral data such as NIR, by using the average spectrum. It does not take any parameters. FT-NIR spectra of Multiplicative Scatter Correction applied corn data is given in Figure 5.7.

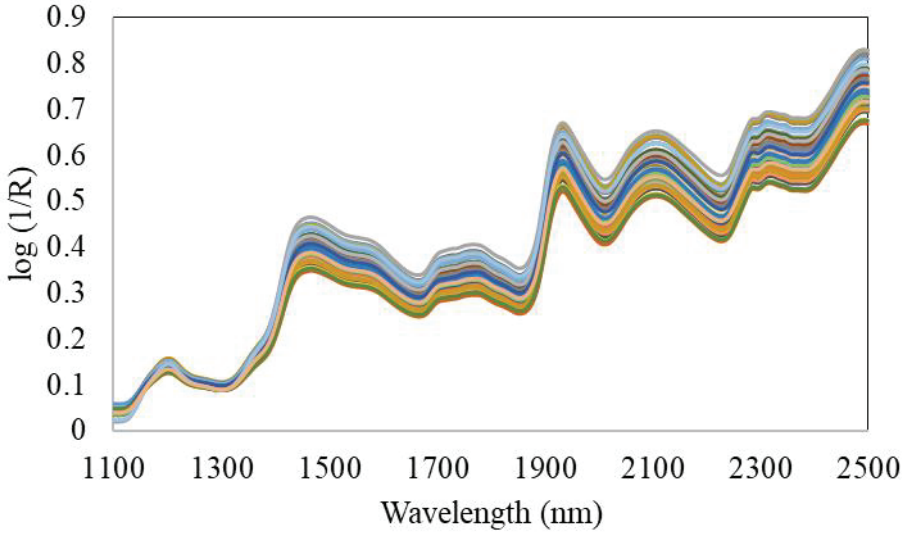


Figure 5.7. FT-NIR spectra of Multiplicative Scatter Correction applied corn data.

In addition to the scattering correction provided by MSC, EMSC is a preprocessing method in which this baseline correction is also performed by ground-fitting the entire spectrum at the same time. Figure 5.8. demonstrates the application of EMSC to the corn data. The spectral features of the corn are conserved, while background offsets and slopes are largely removed compare with Figure 5.2.

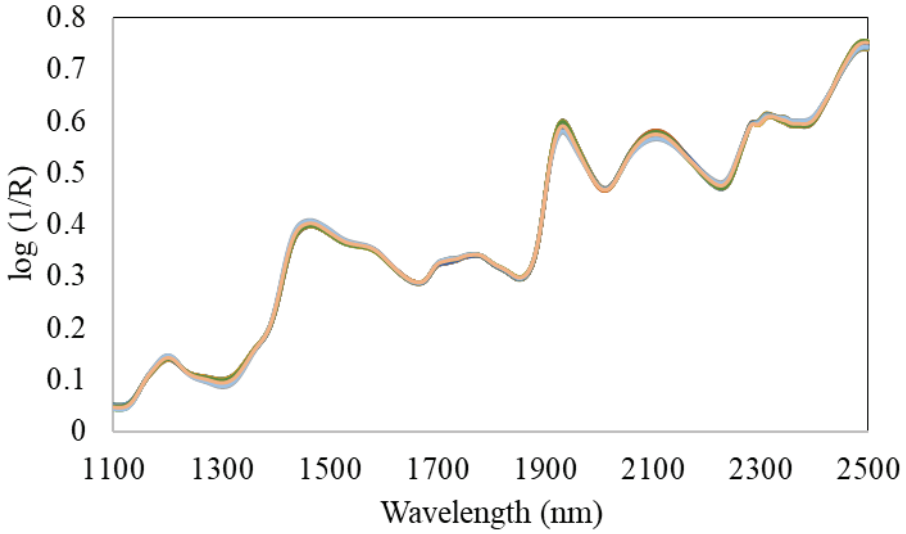


Figure 5.8. FT-NIR spectra of Extended Multiplicative Scatter Correction applied corn data.

Savitzky Golay is used for data smoothing and differentiation. The "Window Size" is used to set the number of points to go in each scroll in the convolution of the filter. It can take the smallest value of 3 and larger numbers can be used for more aggressive smoothing. It must be odd and integer. Besides, this number must be greater than the "Polynomial Order". The Polynomial Order determines the degree of polynomial fit to each window in the filter. It can take at least value of 2 and must be integers. Smaller numbers can be used for a looser fitting, while larger numbers can be used to fit closer to the original data. The “Derivative Order” determines the degree of derivative to be taken. It takes a value of at least 0 and must be chosen as an integer which is less than the degree of the polynomial. Visualization of FT-NIR spectra of Savitzky Golay with using window size of 5; 3rd polynomial order; 1st derivative applied corn data is given in Figure 5.9. and FT-NIR spectra of Savitzky Golay with window size of 5; 3rd polynomial order; 2nd derivative applied corn data is given in Figure 5.10.

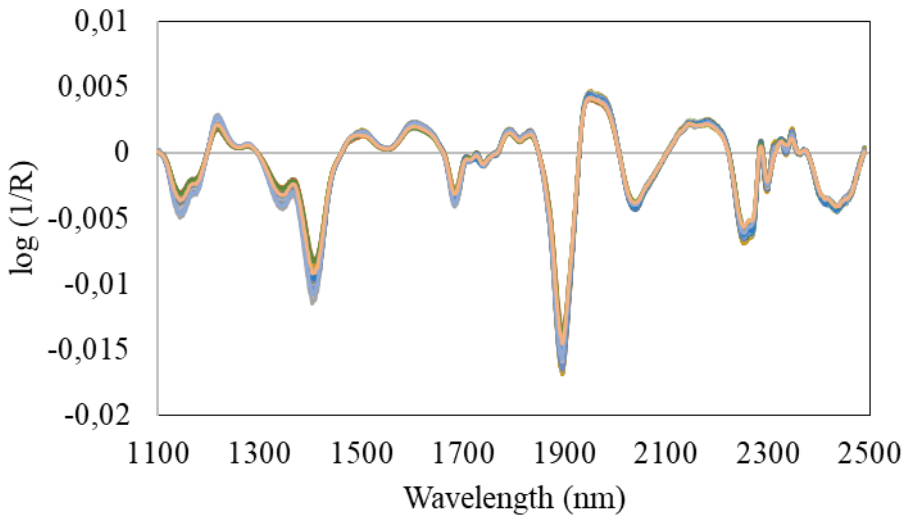


Figure 5.9. FT-NIR spectra of Savitzky Golay with window size of 5; 3rd polynomial order; 1st derivative applied corn data.

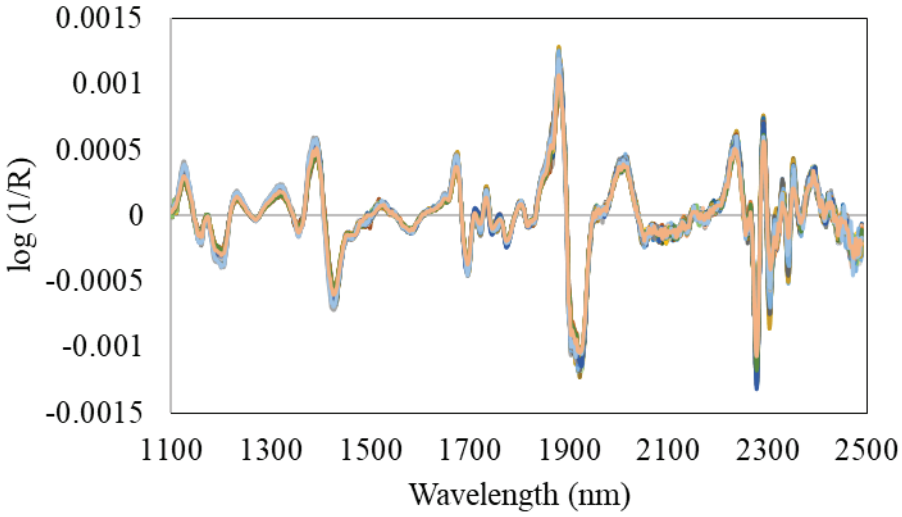


Figure 5.10. FT-NIR spectra of Savitzky Golay with window size of 5; 3rd polynomial order; 2nd derivative applied corn data.

CHAPTER 6

RESULT AND DISCUSSION

6.1. Partial Least Squares Results

Partial Least Squares are performed to observe the effects of applied pretreatment techniques. Results of Partial Least Squares are divided into two categories as PLS results of combinations of different pre-processing techniques on raw data and PLS results of combinations of different pre-processing techniques and mean centering applied to data.

6.1.1. PLS Results of Combinations of Different Pre-processing Techniques on Raw Data

Partial Least Squares is applied on both raw data and combinations of different preprocessing techniques applied data. These preprocessing techniques are Baseline Correction with 1st polynomial order, Baseline Correction with 2nd polynomial order, Standard Normal Variate, Multiplicative Scatter Correction, Extended Multiplicative Scatter Correction, Savitzky Golay with window size of 5; 3rd polynomial order; 1st and 2nd derivative, respectively. Combinations of preprocessing techniques are 1st polynomial order of Baseline Correction plus Savitzky Golay with window size of 5; 3rd polynomial order; 1st and 2nd derivative, 2nd polynomial order of Baseline Correction plus Savitzky Golay with window size of 5; 3rd polynomial order; 1st and 2nd derivative, Standard Normal Variate plus Savitzky Golay with window size of 5; 3rd polynomial order; 1st and 2nd derivative, respectively. PLS results of combinations of different pre-processing techniques on raw data are given in terms of LVs, R^2, SECV and SEP in Table 6.1.

Table 6.1. PLS results of combinations of different pre-processing techniques on raw data.

	LVs	R^2	SECV (w/w %)	SEP (w/w %)
Raw Data	11	0.876	0.181	0.211
BC-1	10	0.877	0.18	0.232
BC-2	9	0.867	0.187	0.236
SNV	11	0.968	0.091	0.105
MSC	11	0.879	0.178	0.222
EMSC	11	0.971	0.085	0.108
SG-5-3-1	6	0.909	0.155	0.27
SG-5-3-2	5	0.924	0.141	0.345
BC-1+SG-5-3-1	6	0.91	0.154	0.265
BC-1+SG-5-3-2	5	0.924	0.141	0.345
BC-2+SG-5-3-1	6	0.913	0.152	0.27
BC-2+SG-5-3-2	5	0.924	0.141	0.345
SNV+SG-5-3-1	6	0.971	0.086	0.138
SNV+SG-5-3-2	5	0.966	0.093	0.222

Initially PLS is used to evaluate Raw Data. For finding the best fitting number of LVs, predicted residual error sum of squares (PRESS) values were calculated for the first 30 LVs and the results are given in Figure 6.1. By using Figure 6.1. PLS models with 11 LVs selected. The R^2 value for calibration set predictions are calculated as 0.876 given in Figure 6.2 (a). As seen on Table 6.1., SECV and SEP values are found to be 0.181 and 0.211, respectively. In order to demonstrate performance of different preprocessing techniques and their combinations on the results of PLS prediction model on corn data, techniques mentioned in the Table 6.1. was applied on Raw Data.

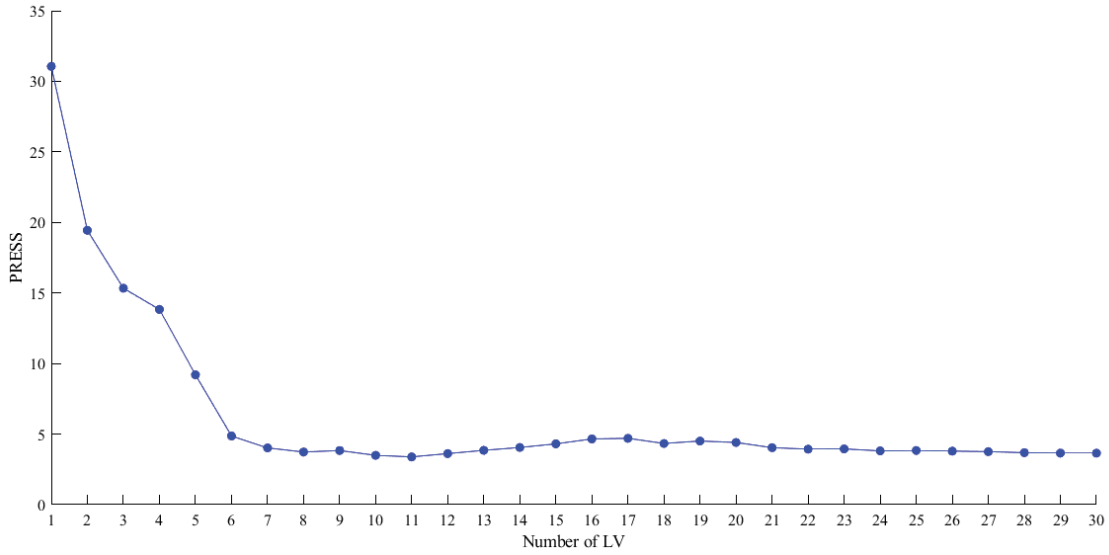


Figure 6.1. Number of LVs vs PRESS plot for Raw Data.

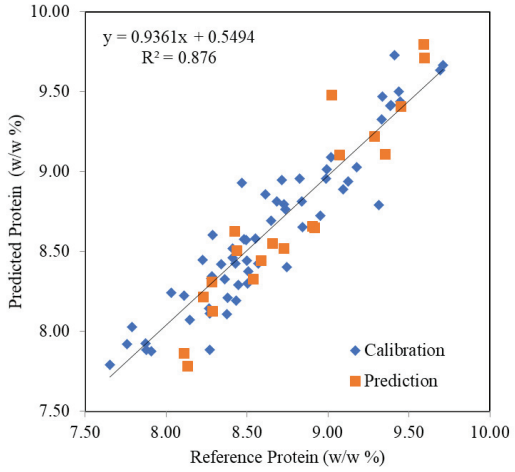
With the purpose of finding best combination of preprocessing, PLS model with 10 and 9 of LVs are performed on BC-1 and BC-2 applied Raw Data, respectively. While SECV and SEP values of BC-1 applied data are 0.180 (w/w %) and 0.232 (w/w %), shown in 6.2 (c), SECV and SEP values of BC-2 applied data are 0.187 (w/w %) and 0.236 (w/w %), shown in 6.2 (d). R^2 results are 0.877 and 0.867, which shows that there is almost no effect of BC-1 and BC-2 preprocessing methods on corn data when compared with the results of PLS model on Raw Data.

Another popular preprocessing method Savitzky Golay is applied and PLS model developed with 6 and 5 of LVs for SG-5-3-1 and SG-5-3-2, respectively. It is observed that even modelling with the relatively low number of components, the model shows better performance. R^2 values of SG-5-3-1 and SG-5-3-2 are 0.909 and 0.924, given in Figure 6.2 (e) and (f) respectively. These values are better in a comparison with R^2 values of raw data, BC-1 and BC-2 applied data. While SECV values, as 0.155 (w/w %), 0.141 (w/w %) for SG-5-3-1 and SG-5-3-2, show decreasement, there are increasement on SEP values, as 0.270 (w/w %), 0.345 (w/w %) for SG-5-3-1 and SG-5-3-2, compared with result of raw data, BC-1 and BC-2 applied corn data. Better investigation of the model, combinations of BC-1+SG-5-3-1, BC-1+SG-5-3-2, BC-2+SG-5-3-1 and BC-2+SG-5-3-2 are performed on data. Apparently, results show that combination of baseline correction and Savitky Golay, no matter which filter chosen, could not yield with better performance

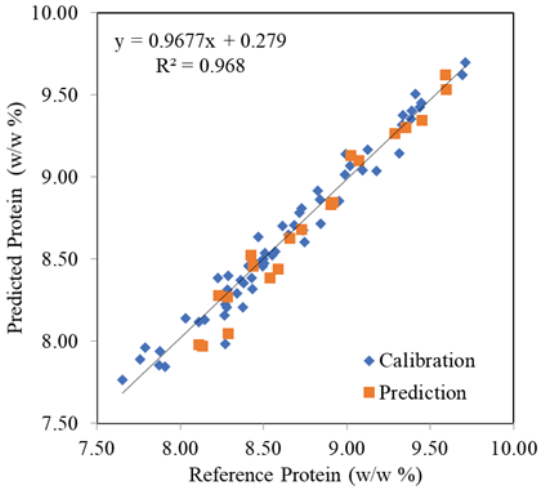
of PLS model when compared with only Savitzky Golay applied raw data. Besides BC-1, BC-2 and their combinations with SG-5-3-1 and SG-5-3-2, SNV and its' combinations with SG-5-3-1 and SG-5-3-2 are carried out. While PLS model developed with 11 LVs for SNV applied data, LVs are 6 and 5 for combinations of SNV+SG-5-3-1 and SNV+ SG-5-3-2 applied data, respectively. Even LVs of PLS model are relatively low for SNV+SG-5-3-1 and SNV+ SG-5-3-2 applied data, R^2 values are 0.971 and 0.966, as given in Figure 6.2 (k) and (l), which are quite close to R^2 value of SNV. SECV values are almost same as 0.091 (w/w %), 0.086 (w/w %) and 0.093 (w/w %) for SNV, SNV+SG-5-3-1 and SNV+ SG-5-3-2 results. However, there is dramatic change on SEP values as while SNV with SEP value of 0.105 (w/w %), SNV+SG-5-3-1 and SNV+ SG-5-3-2 with SEP values of 0.138 (w/w %) and 0.222 (w/w %), given above in Table 6.1.

On the other hand, as known as scatter correction methods, MSC and EMSC are applied to Raw Data. For both cases, 11 of LVs chosen for PLS model. Figure 6.2 (m) gives the R^2 value for MSC applied data is 0.879 which almost same value with results of PLS model on Raw data. There is also not that much difference between SECV values of MSC applied data which is 0.178 (w/w %) and Raw Data. Even there is slightly increasement seen on SEP value of MSC applied data with being 0.222 (w/w %) according to PLS results of Raw Data.

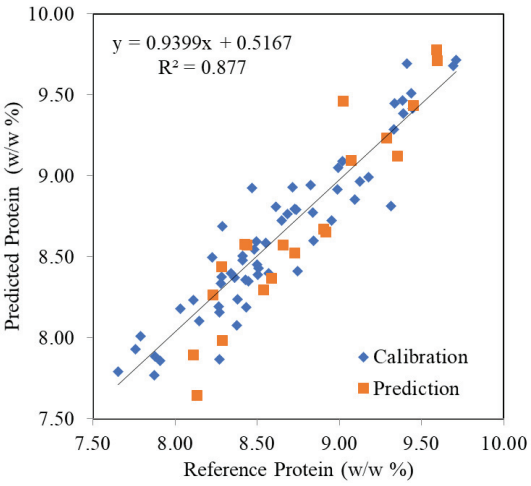
Then again, LVs are choosen as 11 for developement of PLS model of EMSC applied data. Number of LVs vs PRESS plot is given in Figure 6.3 as shown in below. Figure 6.2 (n) shows the R^2 value of model is increased to value of 0.971 by applying EMSC to raw data before carried out PLS prediction. The SECV result is decreased from value of 0.181 (w/w %) for raw data to value of 0.085 (w/w %) for EMSC applied data. The SEP value is also shows dramatic decrease as value of from value of 0.211 (w/w %) for raw data to value of 0.108 (w/w %) for EMSC applied data. The equation of developed PLS model is on Figure 6.4 which shows the reference vs predicted protein (w/w%) graph of EMSC applied data, shown in below.


y = 0.9361x + 0.5494
R² = 0.876
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

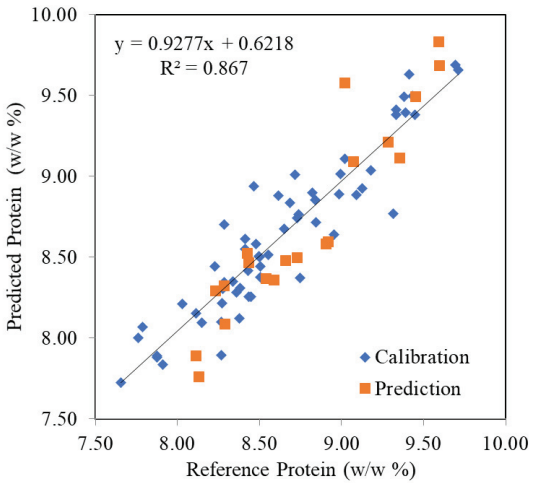
(a)


y = 0.9677x + 0.279
R² = 0.968
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

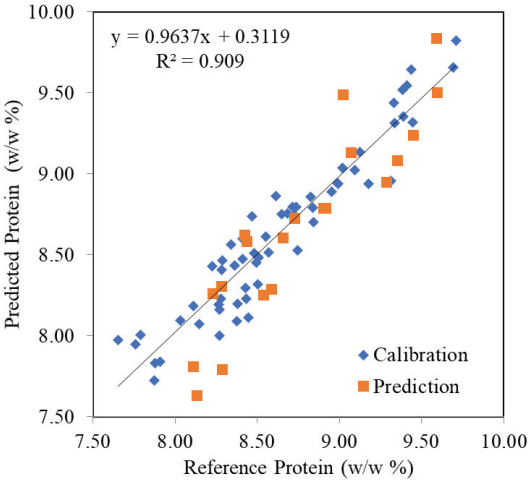
(b)


y = 0.9399x + 0.5167
R² = 0.877
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

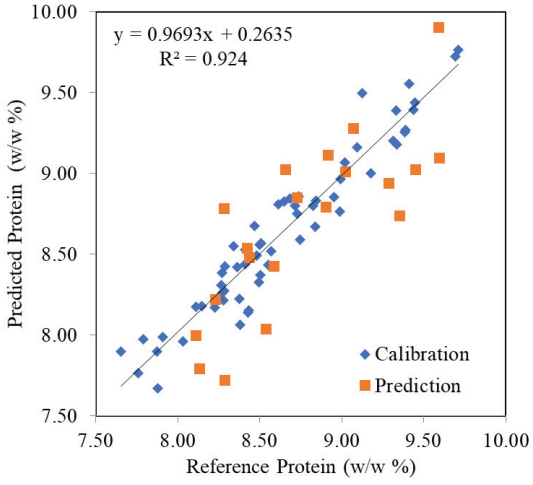
(c)


y = 0.9277x + 0.6218
R² = 0.867
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

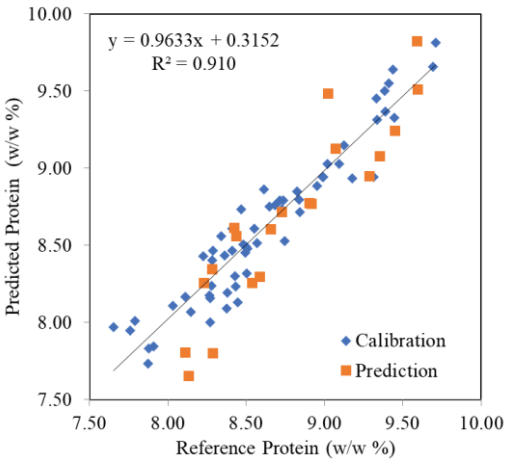
(d)


y = 0.9637x + 0.3119
R² = 0.909
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

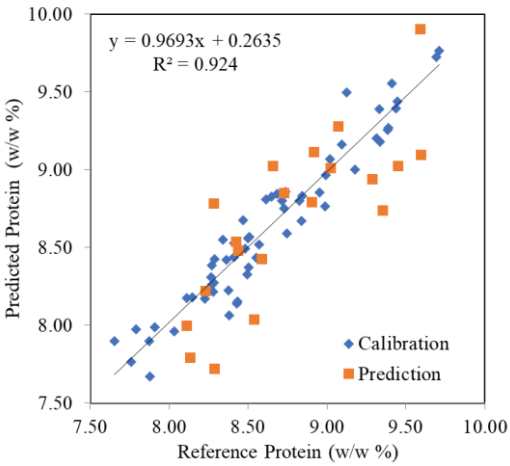
(e)


y = 0.9693x + 0.2635
R² = 0.924
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

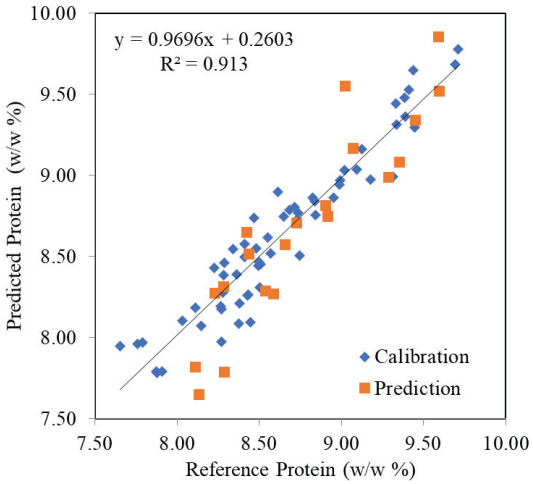
(f)


y = 0.9633x + 0.3152
R² = 0.910
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

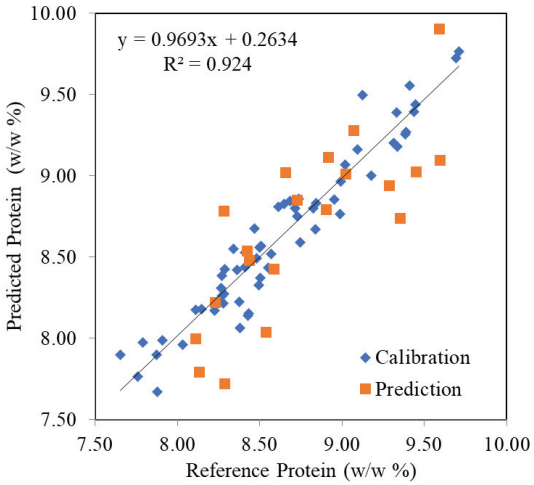
(g)


y = 0.9693x + 0.2635
R² = 0.924
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

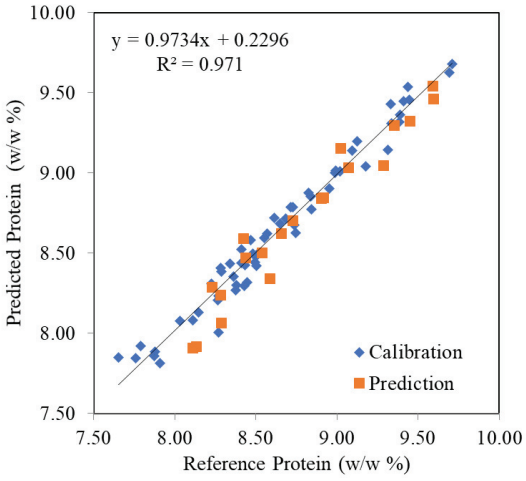
(h)


y = 0.9696x + 0.2603
R² = 0.913
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

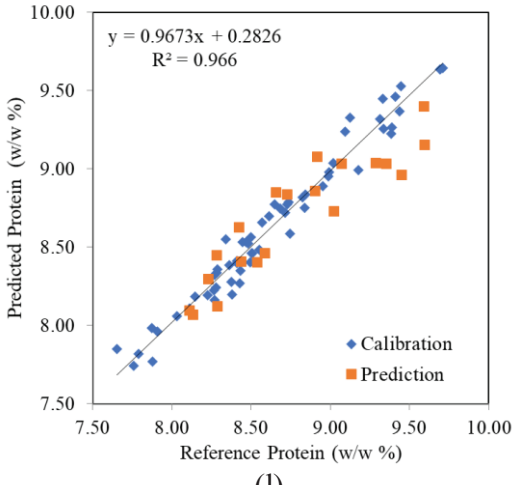
(i)


y = 0.9693x + 0.2634
R² = 0.924
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

(j)


y = 0.9734x + 0.2296
R² = 0.971
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

(k)


y = 0.9673x + 0.2826
R² = 0.966
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

(l)

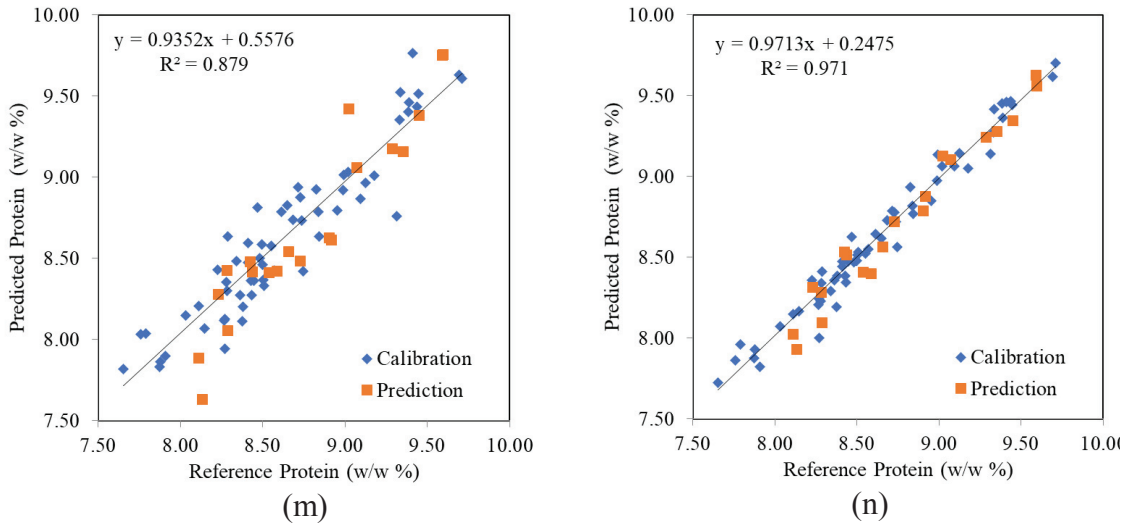


Figure 6.2. Actual concentrations vs. PLS predicted concentrations of protein on a) Raw Data b) SNV c) BC-1 d) BC-2 e) SG-5-3-1 f) SG-5-3-2 g) BC-1+SG-5-3-1 h) BC-1+SG-5-3-2 i) BC-2+SG-5-3-1 j) BC-2+SG-5-3-2 k) SNV+SG-5-3-1 l) SNV+SG-5-3-2 m) MSC n) EMSC.

6.1.2. PLS Results of Combinations of Different Pre-processing Techniques and Mean Centering Applied to Data

Partial Least Squares is applied on combinations of different preprocessing techniques plus mean centering applied data. The order of the applied preprocessing techniques and their combinations are as same as descried in section 6.1.1. PLS results of combinations of different pre-processing techniques on mean centering applied data are given in terms of LVs, R^2, SECV and SEP in Table 6.2.

Table 6.2. PLS results of combinations of different pre-processing techniques on mean centering applied data.

	LVs	R^2	SECV (w/w %)	SEP (w/w %)
MC	19	0.995	0.035	0.08
BC-1+ MC	16	0.994	0.038	0.088
BC-2+MC	16	0.995	0.037	0.083
SNV+MC	12	0.977	0.076	0.112
MSC+MC	17	0.993	0.043	0.094
EMSC+MC	10	0.971	0.085	0.108
SG-5-3-1+MC	7	0.992	0.044	0.121
SG-5-3-2+MC	5	0.984	0.064	0.207
BC-1+SG-5-3-1+MC	7	0.99	0.051	0.115
BC-1+SG-5-3-2+MC	7	0.996	0.033	0.203
BC-2+SG-5-3-1+MC	8	0.992	0.045	0.116
BC-2+SG-5-3-2+MC	7	0.996	0.033	0.203
SNV+SG-5-3-1+MC	6	0.975	0.08	0.145
SNV+SG-5-3-2+MC	6	0.993	0.044	0.186

Mean centering is applied in order to have better axis fitting to the set zero mean value. To be able to observe the effect of mean centering technique on corn data, PLS model developed with 19 of LVs, as shown in Figure 6.3.

As shown in Figure 6.4 (a), R^2 value of the new developed method is 0.995 which is showed significant increase. In the Table 6.2, SECV (w/w %) and SEP are 0.035 (w/w %) and 0.080 that error values show that application of mean centering to the raw data makes model of corn data more evaluated.

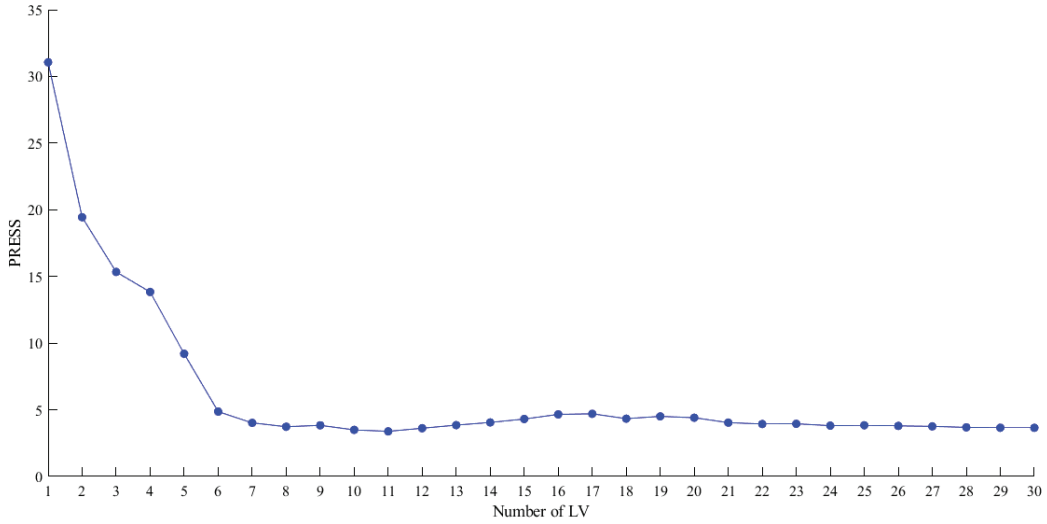


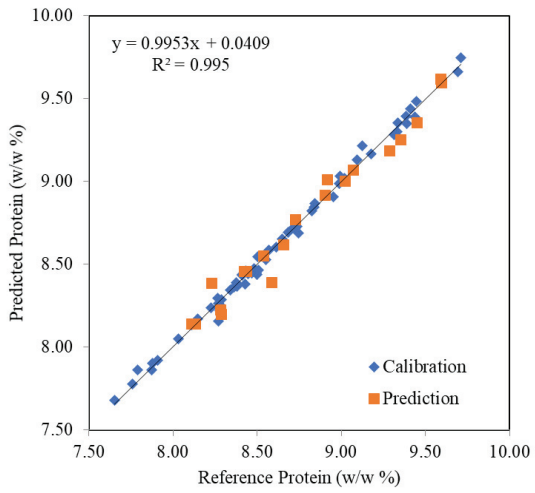
Figure 6.3. Number of LVs vs PRESS plot for MC applied data.

Afterwards BC-1 and BC-2 performed, and MC applied to the data. In the Figure 6.4 (c) and (d), R^2 results of model are 0.994 and 0.995 while LVs are 16 for both BC-1+MC and BC-2+MC. SECV and SEP results are 0.038 (w/w %),0.037 (w/w %) and 0.088 (w/w %), 0.083 (w/w %) for BC-1+MC and BC-2+MC respectively in Table 6.2. Results demonstrate that efficiency of model is practically the same with slightly lower latent variables compared with model which is developed on only MC applied data. Furthermore SG-5-3-1 and SG-5-3-2 are performed, and MC applied to corn data. Even with a few latent variables which are 7 and 5 for SG-5-3-1+MC and SG-5-3-2+MC, R^2 values are still close to 1 as being 0.992 and 0.984 which are given in Figure 6.4 (e) and (f). However, SECV and SEP values 0.044 (w/w %) and 0.121 (w/w %) for SG-5-3-1+MC and 0.064 (w/w %) and 0.207 (w/w %) for SG-5-3-2+MC, which are higher than PLS model reaults of MC performed corn data. Furthermore, new PLS models developed after followed combinations of preprocessing techniques performed on corn data; BC-1+SG-5-3-1+MC, BC-1+SG-5-3-2+MC, BC-2+SG-5-3-1+MC and BC-2+SG-5-3-2+MC. R^2 values are 0.990, 0.996, 0.992 and 0.996 which are shown in Figure 6.4 (g), (h), (i) and (j). LVs of these combinations are 7,7,8 and 7, respectively as given in Table 6.2. After adding SG-5-3-1 to the combinations of BC-1+MC and BC-2+MC; SECV is becoming to 0.051 (w/w %) and 0.045 (w/w %), SEP is becoming to 0.115 (w/w %) and 0.116 (w/w %) as shown in Table 6.2. As regards results of PLS model on combination of BC-1+MC and BC-2+MC applied data, error valuse are inreasising by adding SG-5-3-1. Besides while adding SG-5-3-2 to BC-1+MC and BC-2+MC make SECV results

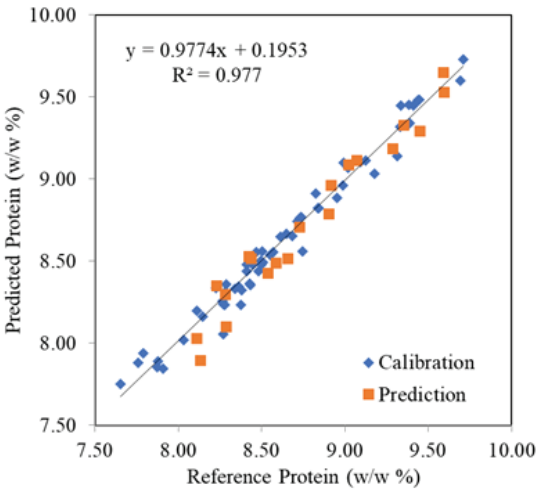
decrease, increasing on SEP results observed. SECV and SEP value for 0.033 and 0.203 for both combinations of BC-1+SG-5-3-2+MC and BC-2+SG-5-3-2+MC. The reason of increasing on errors can be explain as reducing necessary information on data while performing combinations of preprocessing.

Another preprocessing technique of SNV performed then MC applied on data. The PLS model evaluated with 12 of LVs. The R^2 value is 0.977 as given in Figure 6.4 (b). Additionally, SECV and SEP results are 0.076 (w/w %) and 0.112 (w/w %). While PLS model is developed in a better way after applying SNV+MC according to the raw data, results show ruining when compared with PLS model of only MC applied data. Besides, SNV+SG-5-3-1+MC and SNV+SG-5-3-2+MC combinations have been performed. Both PLS models are developed with selecting 6 LVs. For the PLS model with combination of SNV+SG-5-3-1+MC R^2 value is 0.975 as shown in Figure 6.4 (k). SECV and SEP results are 0.080 and 0.145. For the combination of SNV+SG-5-3-2+MC, R^2 value is 0.993 as shown in Figure 6.4 (l). SECV result is 0.044 (w/w %) but SEP result is increasing from 0.080 (w/w %) to 0.186 (w/w %) according to PLS model results of MC applied data.

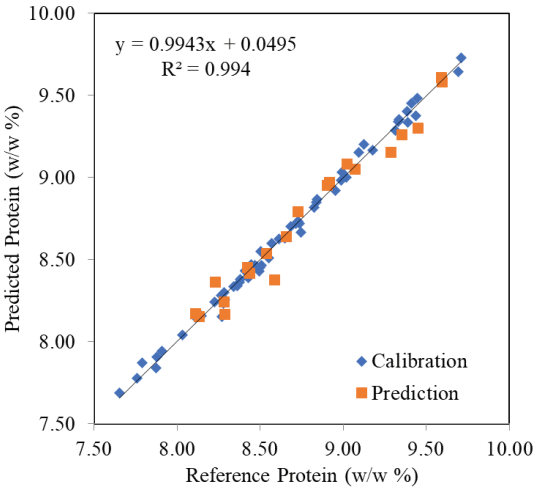
Later on, another PLS models developed with LVs of 17 and 10 respectively for MSC+MC and EMSC+MC combinations applied data. The effect of selection of LVs can be observed on R^2 results as being 0.993 and 0.971 for MSC+MC and EMSC+MC combinations which are shown with their equations in Figure 6.4 (m) and (n). The SECV and SEP results of PLS model on corn data treated with combinations of MSC+MC and EMSC+MC are 0.043 (w/w %),0.094 (w/w %) and 0.085 (w/w %), 0.108 (w/w %).


y = 0.9953x + 0.0409
R² = 0.995
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

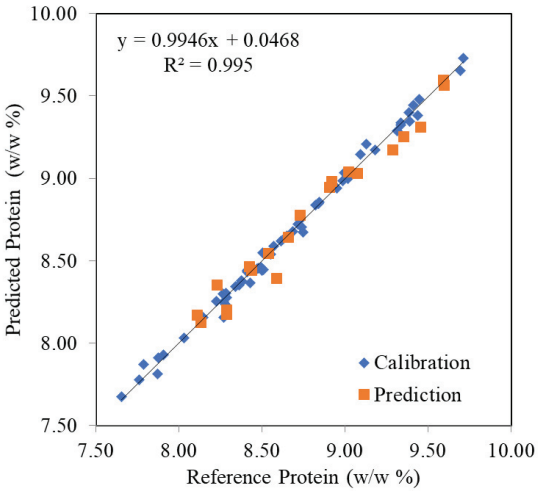
(a)


y = 0.9774x + 0.1953
R² = 0.977
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

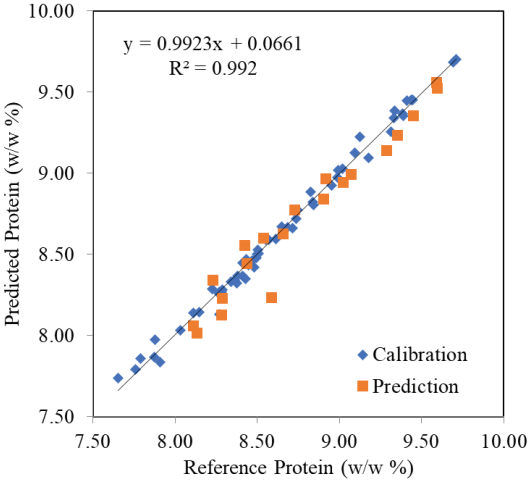
(b)


y = 0.9943x + 0.0495
R² = 0.994
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

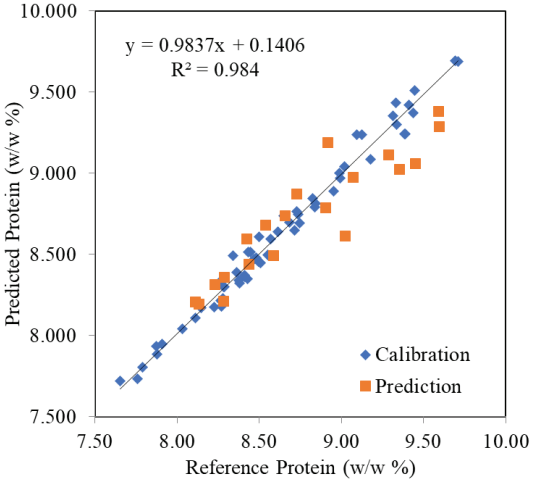
(c)


y = 0.9946x + 0.0468
R² = 0.995
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

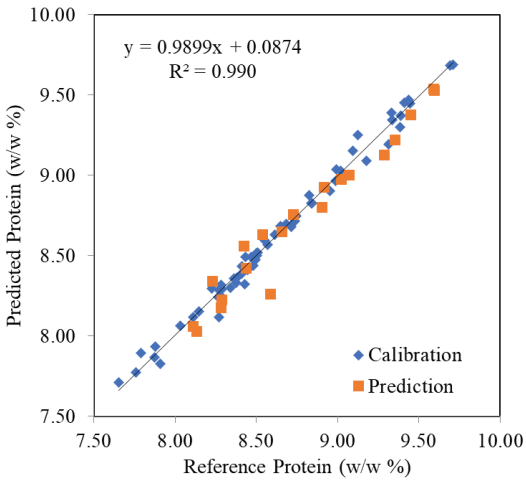
(d)


y = 0.9923x + 0.0661
R² = 0.992
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

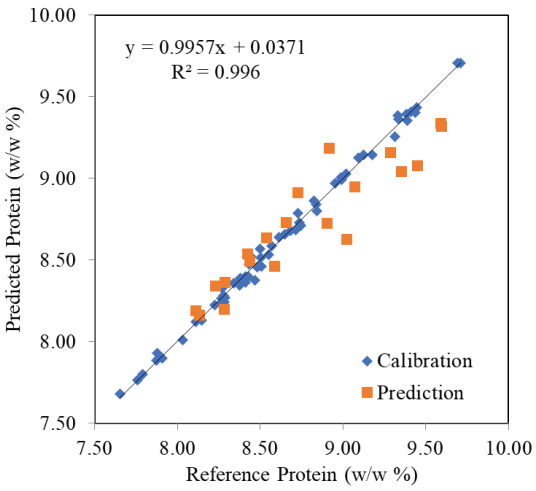
(e)


y = 0.9837x + 0.1406
R² = 0.984
Predicted Protein (w/w %)
Reference Protein (w/w %)
Calibration
Prediction
10.000
9.500
9.000
8.500
8.000
7.500
7.50
8.00
8.50
9.00
9.50
10.00

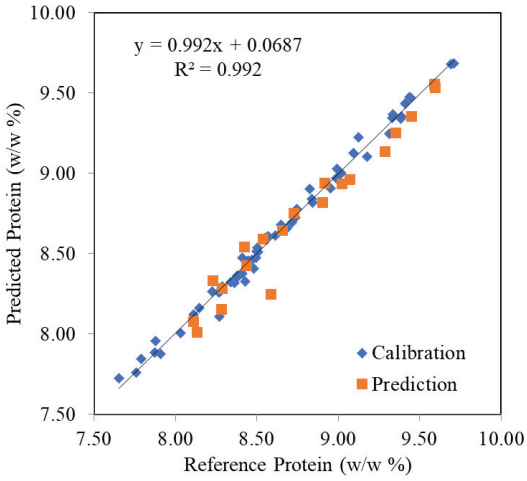
(f)


y = 0.9899x + 0.0874
R² = 0.990
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

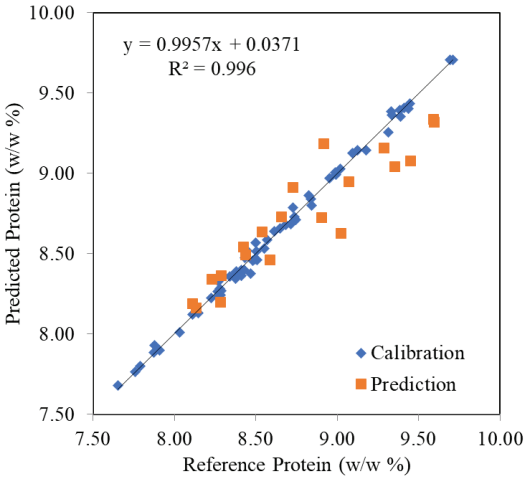
(g)


y = 0.9957x + 0.0371
R² = 0.996
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

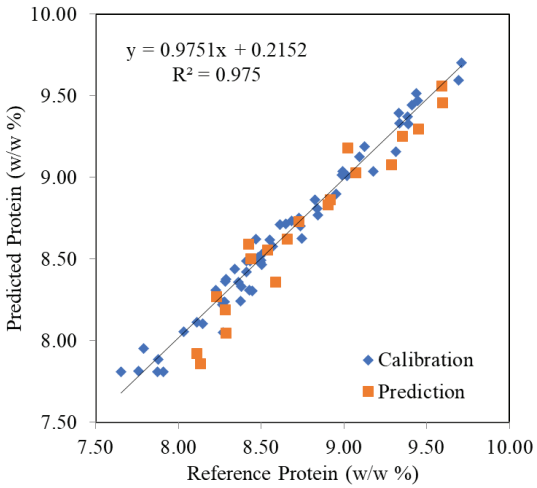
(h)


y = 0.992x + 0.0687
R² = 0.992
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

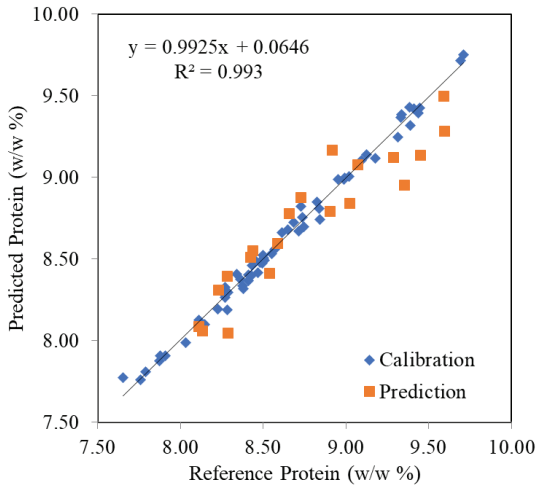
(i)


y = 0.9957x + 0.0371
R² = 0.996
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

(j)


y = 0.9751x + 0.2152
R² = 0.975
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

(k)


y = 0.9925x + 0.0646
R² = 0.993
Calibration
Prediction
Predicted Protein (w/w %)
Reference Protein (w/w %)
10.00
9.50
9.00
8.50
8.00
7.50
7.50
8.00
8.50
9.00
9.50
10.00

(l)

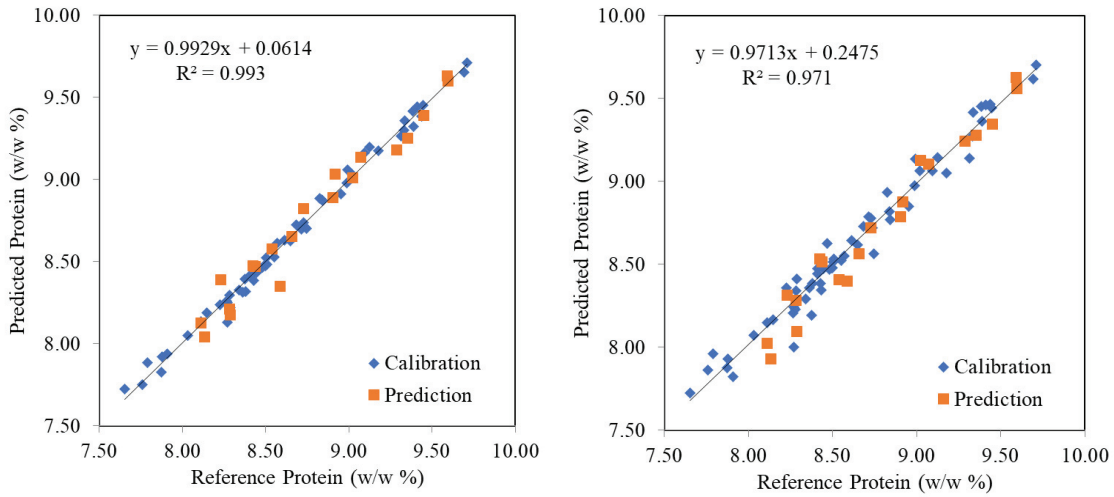


Figure 6.4. Actual concentrations vs. PLS predicted concentrations of protein on a) MC b) SNV+MC c) BC-1+MC d) BC-2+MC e) SG-5-3-1+MC f) SG-5-3-2+MC g) BC-1+SG-5-3-1+MC h) BC-1+SG-5-3-2+MC i) BC-2+SG-5-3-1+MC j) BC-2+SG-5-3-2+MC k) SNV+SG-5-3-1+MC l) SNV+SG-5-3-2+MC m) MSC+MC n) EMSC+MC

6.2. Genetic Inverse Least Squares Results

Another one of the most popular multivariate calibration method Genetic Inverse Least Square (GILS) developed to observe whether the approximation on model have effect or not. Since mean centering have been observed as the most effective preprocessing technique on protein component of FT-NIR corn data, MC also applied before evaluate GILS model to compare results fairly. After MC pre-processing performed on raw NIR spectra, GILS with 30 genes, 50 iterations and 100 runs where R^2 threshold for selection of initial genes were 0.5 and 1-fold CV was used for determination of fitness was applied to establish prediction models.

The Figure 6.5 demonstrates selection frequencies of gens and indicates the range with the most selected genes. Since threshold is define as 0.5, this graph shows how many genes with R^2 higher than 0.5 selected to the pool. As seen on the graph especially in the

variable index of between 500 and 550, there are more genes selected. It explains that there is significant information observed for the model in this range.

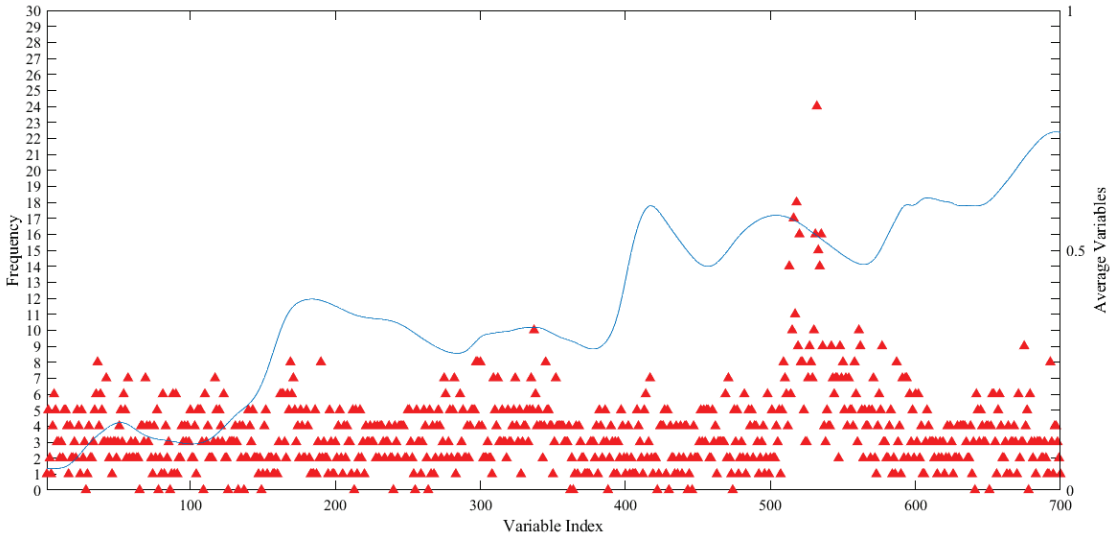


Figure 6.5. Graph of Variables vs. Selection Frequencies.

The success of model performance can be assessable with R^2 value of 0.997 as given in figure 6.6. SECV and SEP results of GILS model are 0.025 (w/w %) and 0.042 (w/w %) respectively.

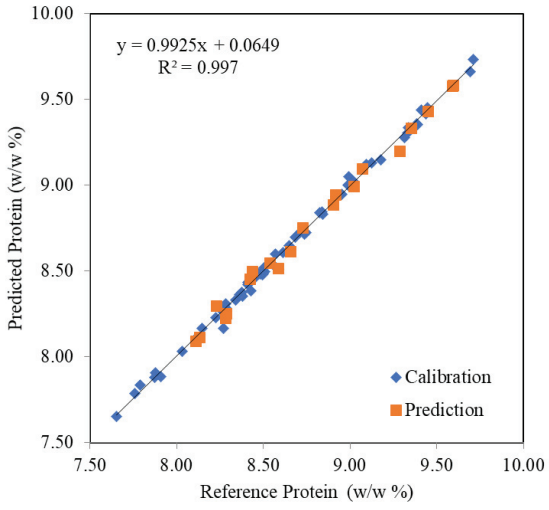


Figure 6.6. Actual concentrations vs. PLS predicted concentrations of GILS model on MC applied data.

6.3. Summary and Comparison of PLS and GILS Results

In the scope of the thesis, various preprocessing techniques and their combinations performed on protein component of FT-NIR corn data. PLS and GILS models developed to be able to show the effect of techniques. Reference protein (w/w%) values and predicted protein (w/w%) values of PLS and GILS results of mean centering applied data are given in Table 6.3.

Table 6.3. Reference protein (w/w%) values and predicted protein (w/w%) values of PLS and GILS results of mean centering applied data.

Reference Protein (w/w%)	**PLS results' Predicted Protein (w/w%)**	**GILS results' Predicted Protein (w/w%)**
8.658	8.516	8.611
9.595	9.528	9.579
8.918	8.960	8.943

Comparison of PLS results in the manner of SEP and SECV results are given in Figure 6.7 and 6.8. On those figures, dark column represents error results of combinations of preprocessing techniques applied raw data which is also named as without mean centering. Light column represents error results of after application of those combinations of preprocessing techniques mean centering performed data which is also named as with mean centering.

Figure 6.7 demonstrate that for most the combinations, prediction errors of raw data are higher according to mean centered data. Error values are same only for EMSC applied raw data and mean centered data. Exceptions are SNV and SNV+SG-5-3-1 for these cases with being higher SEP results with just barely different. It can be observed from figure that the least error of prediction is the result of PLS model with only mean centered data. The highest errors of prediction results are for both combinations with SG-5-3-2. It explains that application of 2nd derivative Savitzky Golay filter have reverse effect on prediction errors.

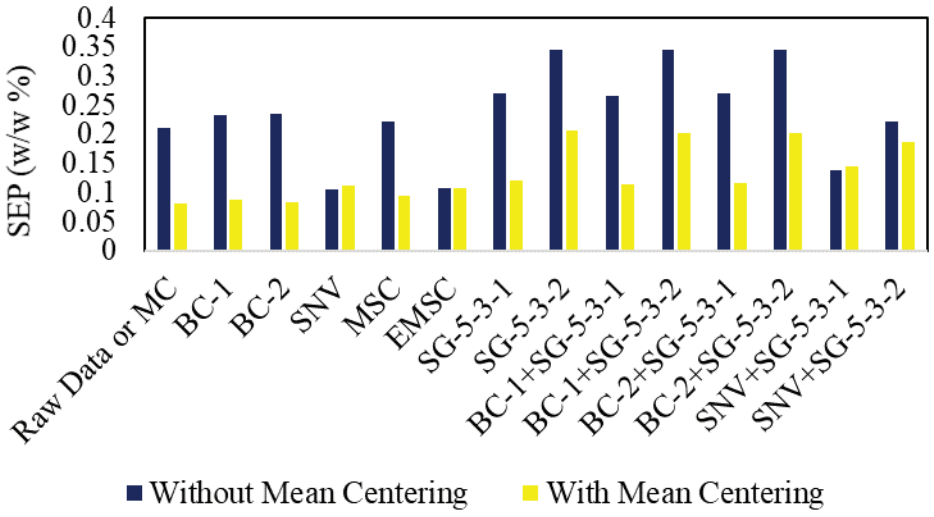


Figure 6.7. Comparison of SEP results of raw data and mean centered data.

As can be seen on Figure 6.8, SEP results of raw data are again higher then mean centered data except only EMSC. The error of calibration for EMCS applied data are same for both raw and mean centering applied data. 2nd polynomial order baseline correction applied raw data shows the highest error of calibration. Among all the combinations the lowest SECV result is for only mean centering applied data.

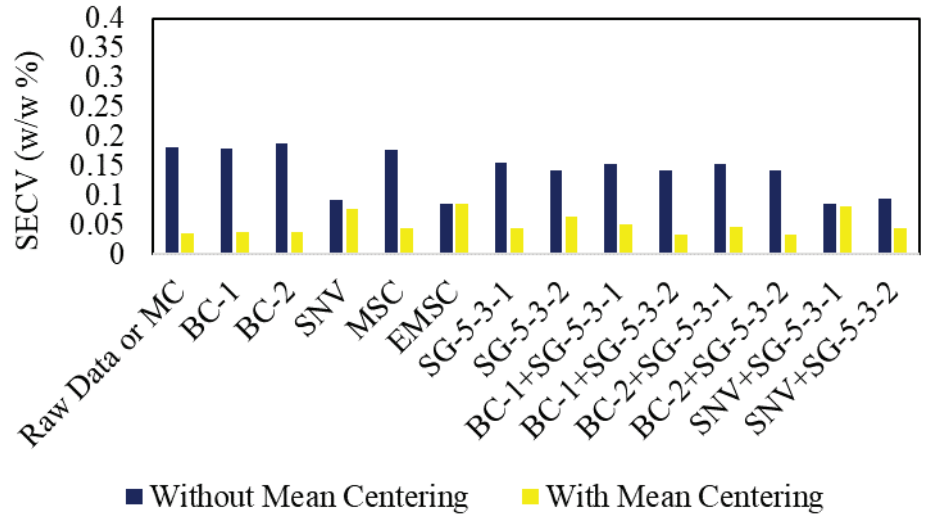


Figure 6.8. Comparison of SECV results of raw data and mean centered data.

Since both SECV and SEP results of mean centered data are lowest, GILS model developed with mean centering performed data. SECV and SEP results of PLS model is 0.035 (w/w %) and 0.080 (w/w %). For GILS model, SECV and SEP results are 0.025 (w/w %) and 0.042 (w/w %) which are lower compared with results of PLS model. Furthermore, while the R^2 value of PLS model is 0.995, it is 0.997 for GILS model.

Better in comparison, t-Test: Paired Two Sample for Means applied between those reference corn data and predicted protein of raw data, reference protein and predicted protein of mean centered data, predicted protein of raw data and predicted corn data of mean centered data. The p values are 0.195, 0.410 and 0.345 respectively. The results show there is no significant difference within 95% confidence level between the performance of the PLS model without preprocessing and PLS model of mean centered data. When t-Test: Paired Two Sample for Means applied to reference corn data and predicted protein of mean centered data of GILS model. The p value result is 0.223 and demonstrates there is no significant difference within 95% confidence level.

CHAPTER 7

CONCLUSION

In this thesis, several pre-processing techniques and different combinations of them have been processed on corn NIR data. Their effects are examined by comparing the results of two different multivariate calibration methods which are GILS and PLS.

Among the carried preprocessing techniques which are mean centering, spectral derivatives and scatter-corrections and all combinations of them, the GILS and PLS of mean centering applied corn NIR data are results with lowest SEP value which are 0.042 (w/w %) and 0.080 (w/w %), respectively.

The p values of t-test results of GILS and PLS between the reference and predicted protein of mean centered data are 0.223 and 0.410, respectively. These p value results demonstrate that there is no significant difference within 95% confidence level. Obviously, quantitative results do not give suggestion about which pre-processing technique to use in any case for corn NIR data set. However, it does appear sensible to applying mean centering to the corn NIR data makes SEP value more decrease while evaluating performance of PLS model comparing with other preprocessing techniques or their combinations. Besides, the biggest decrease on SEP value, which is good manner, observed from GILS model with mean centered data.

REFERENCES

Afseth, N. K., & Kohler, A. (2012). Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems*, *117*, 92-99.

Akkoç, G. D., Development of chemometrics calibration toolbox and its application for determination of slep adulteration. Izmir Institute of Technology, 2018.

Andrew, A., & Fearn, T. (2004). Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. *Chemometrics and Intelligent Laboratory Systems*, *72*(1), 51-56.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1994). Correction to the description of standard normal variate (SNV) and de-trend (DT) transformations in Practical Spectroscopy with Applications in Food and Beverage Analysis—2nd edition. *NIR news*, *5*(3), 6-6.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, *43*(5), 772-777.

Başar, B., Development of fast and simple analytical methods for the determination of honey adulteration and forgery based on chemometric multivariate data analysis by using molecular spectroscopy. Izmir Institute of Technology, 2016.

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

Dai, S., Pan, X., Ma, L., Huang, X., Du, C., Qiao, Y., & Wu, Z. (2018). Discovery of the linear region of near infrared diffuse reflectance spectra using the kubelka-munk theory. *Frontiers in chemistry*, *6*, 154.

Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, *50*, 96-106.

Eilers, P. H. (2004). Parametric time warping. *Analytical chemistry*, *76*(2), 404-411.

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied spectroscopy*, *39*(3), 491-500.

Geladi, Paul, and Bruce R Kowalski. (1986). "Partial least-squares regression: a tutorial." *Analytica chimica acta* 185:1-17.

Isaksson, T., & Næs, T. (1988). The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *Applied Spectroscopy*, *42*(7), 1273-1284.

Kawamura, K., Tsujimoto, Y., Nishigaki, T., Andriamananjara, A., Rabenarivo, M., Asai, H., ... & Razafimbelo, T. (2019). Laboratory visible and near-infrared spectroscopy with genetic algorithm-based partial least squares regression for assessing the soil phosphorus content of upland and lowland rice fields in Madagascar. *Remote Sensing*, *11*(5), 506.

Laxalde, J., Ruckebusch, C., Devos, O., Caillol, N., Wahl, F., & Duponchel, L. (2011). Characterisation of heavy oils using near-infrared spectroscopy: optimisation of pre-processing methods and variable selection. *Analytica Chimica Acta*, *705*(1-2), 227-234.

Martens, H., Jensen, S. A., & Geladi, P. (1983, June). Multivariate linearity transformation for near-infrared reflectance spectrometry. In *Proceedings of the Nordic symposium on applied statistics* (pp. 205-234). Stokkand Forlag Publishers Stavanger, Norway.

Meşe, A. E., Development of a new infrared spectroscopic method based on multivariate calibration for the determination of aluminum and magnesium oxide thickness on aluminum foil and sheets surfaces. Izmir Institute of Technology, 2016.

Mishra, P., Roger, J. M., Rutledge, D. N., & Woltering, E. (2020). SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials. *Postharvest Biology and Technology*, *168*, 111271.

Ni, W., Brown, S. D., & Man, R. (2009). Stacked partial least squares regression analysis for spectral calibration and prediction. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *23*(10), 505-517.

Özdemir, D., and Öztürk, B., 2004. "Genetic multivariate calibration methods for near infrared (NIR) spectroscopic determination of complex mixtures." *Turkish Journal of Chemistry* 28 (4):497-514.

Pasquini, C., Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian chemical society* 2003, *14* (2), 198-219.

Rabatel, G., Marini, F., Walczak, B., & Roger, J. M. (2020). VSN: Variable sorting for normalization. *Journal of Chemometrics*, *34*(2), e3164.

Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201-1222.

Roger, J. M., Biancolillo, A., & Marini, F. (2020). Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *199*, 103975.

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

Skoog, Douglas A, F James Holler, and Stanley R Crouch. 2017. *Principles of instrumental analysis.*

Stuart, B., Experimental methods. *Infrared spectroscopy: fundamentals and applications* 2004, 18-19.

Şentürk, S., Determination of hydrocarbon composition of naphtha by using fourier transform infrared spectroscopy and multivariate calibration. Izmir Institute of Technology, 2020.

Tan, H., & Brown, S. D. (2003). Multivariate calibration of spectral data using dual-domain regression analysis. *Analytica chimica acta*, *490*(1-2), 291-301.

URL-1 <http://obakemometri.com>