QUANTITATIVE AND QUALITATIVE ASSESSMENT OF PETROLEUM PRODUCTS BY SPECTROSCOPIC AND CHEMICAL DATA ANALYTICS

ABD RAHIM OTHMAN

FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2022

QUANTITATIVE AND QUALITATIVE ASSESSMENT OF PETROLEUM PRODUCTS BY SPECTROSCOPIC AND CHEMICAL DATA ANALYTICS

ABD RAHIM OTHMAN

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF SCIENCE

DEPARTMENT OF CHEMISTRY
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2022

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **Abd Rahim Bin Othman**

Matric No: **S2006467**

Name of Degree: **MASTER OF SCIENCE**

Title of Dissertation ("this Work"):

QUANTITATIVE AND QUALITATIVE ASSESSMENT OF PETROLEUM PRODUCTS BY SPECTROSCOPIC AND CHEMICAL DATA ANALYTICS

Field of Study:

ANALYTICAL CHEMISTRY & CHEMOMETRIC

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 13th March 2023

Subscribed and solemnly declared before,

Witness's Signature Date: 13th March 2023

QUANTITATIVE AND QUALITATIVE ASSESSMENT OF PETROLEUM PRODUCTS BY SPECTROSCOPIC AND CHEMICAL DATA ANALYTICS

ABSTRACT

This research study describes the use of Fourier Transform Near Infra-Red (FT-NIR) technology with various chemometric methods in petroleum products. Current practice in the refinery's quality assurance and quality control laboratories relies on the use of conventional method measurements in the quantitative and qualitative assessment of their petroleum products. These conventional methods consume high workforce, longer analysis time and high operating expenditure. A new innovative way of qualitative and quantitative measurement and application were explored to improve and address the pain points and provide a total solution. This study aims to evaluate and assess the quality of petroleum products through spectroscopic and chemical data analytics, both qualitative and quantitative assessments. In this work, coupled near infra-red spectroscopy and chemometrics techniques for calibration models development, i.e. Partial Least Square (PLS) and Principal Component Regression (PCR) and Near Infra-red (NIR) spectra pre-treatment or re-processing, i.e. Multiplicative Scatter Correction (MSC) and Savitzky Golay Second Derivative (SGSD) were employed accordingly for rapid and simultaneous determination of chemical and physical properties of petroleum hydrocarbons. The coupled NIR and chemometrics methods are an alternative to the existing laboratory reference methods to address the refinery pain points. FT-NIR spectroscopy has been successfully utilized to rapidly identify and discriminate three types of petroleum products (gasoline, diesel, and kerosene) using Principal Component Analysis (PCA). More than 95% of each product was accurately identified and differentiated. This qualitative multivariate measurement is important when fast results are required at the operation site, such as during product transfer cross-contamination and adulteration or

illegal product blending. In addition, qualitative measurement by PCA was used to differentiate gasoline and diesel fuels directly sourced from refineries without additive. In contrast, additives were added to the gasoline and diesel fuels, such as corrosion inhibitors, detergency, and lubricity improvers, to enhance the engine's performance and protection of the engine components. Diesel with and without palm methyl ester (PME) blend were also determined qualitatively using PCA based on significant the presence of fatty acid methyl ester (FAME) in diesel. This work demonstrates the multivariate calibration strategy for the simultaneous near-infrared spectrometric determination of the physical and chemical properties of the petroleum products, namely the boiling point at 95% recovery (T95%), flash point (FP), cloud point (CP) and cetane index (CI) which include the spectral region selection, calibration/validation set partition, data pre-processing, and regression. Based on the results, the calibration constructed on the combination region of 4800-4000 cm^{-1} using the randomly selected calibration set managed to deliver excellent predictive performance in terms of coefficient of determination, root mean square error of cross-validation, root mean square error of prediction and the ratio of performance deviation. Moreover, all the developed models satisfied the reproducibility requirement of respective American Society for Testing and Materials (ASTM) standard methods regardless of the employment of multiplicative scattering correction/Savitzky-Golay second-derivatization and principal component regression/partial least square regression. This revealed that the fitness of the model relies upon every single calibration component. It was also realized that data pre-treatment is crucial in delivering predictive-performing predictions.

Keywords: qualitative, quantitative, chemometrics, petroleum products

QUANTITATIVE AND QUALITATIVE ASSESSMENT OF PETROLEUM PRODUCTS BY SPECTROSCOPIC AND CHEMICAL DATA ANALYTICS

ABSTRAK

Penyelidikan ini membincangkan penggunaan teknologi *Fourier Transform Near Infra-Red (FT-NIR)* menggunakan pelbagai kaedah kimometrik dalam produk petrolium. Amalan di makmal jaminan kualiti dan kawalan kualiti kilang masa kini, bergantung pada penggunaan pengukuran kaedah konvensional dalam menilai tahap kuantitatif dan kualitatif produk petrolium mereka. Kaedah konvensional ini menggunakan tenaga kerja yang tinggi, masa analisa yang lebih lama dan perbelanjaan operasi yang tinggi. Kaedah inovatif mengukur aplikasi kualitatif dan kuantitatif baru diterokai untuk memperbaiki dan mengatasi masalah ini dan memberikan penyelesaian secara menyeluruh. Kajian ini dijalankan bertujuan untuk mentaksir dan menilai kualiti produk petrolium melalui spektroskopi dan analisis data kimia bagi penilaian kualitatif dan kuantitatif. Dalam hasil kerja ini, teknik spektroskopi infra-merah dan kimometrik untuk pembangunan model penentukuran, seperti sebahagian kecil daripada persegi (PLS) dan regresi komponen utama (PCR) dan pra-rawatan spektrum NIR atau pemprosesan semula, seperti contoh pembetulan penyebaran berbilang (MSC) dan *Savitzky Golay* Kedua Derivatif (SGSD) telah digunapakai untuk penentuan pesat dan serentak, sifat kimia dan fizikal bagi hidrokarbon petrolium. Kaedah NIR dan kimometrik yang digabungkan adalah alternatif kepada kaedah rujukan makmal yang sedia ada untuk mengatasi masalah kilang penapisan. Spektroskopi FT-NIR telah berjaya digunakan untuk mengenal pasti tiga jenis produk petroleum, iaitu petrol, diesel, dan minyak tanah menggunakan analisa komponen utama (PCA). Lebih 95% daripada setiap produk telah dikenalpasti dan dibezakan dengan tepat. Hal ini amat penting apabila keputusan segera diperlukan di tapak operasi, seperti semasa pemindahan produk pencemaran silang dan pengadukan atau pengadunan produk

secara haram. Tambahan pula, pengukuran kualitatif oleh PCA digunakan untuk membezakan bahan bahan api petrol dan diesel secara langsung yang diperoleh dari kilang penapis di mana bebas bahan tambahan dicampur dalam bahan api tersebut. Sebaliknya, bahan tambahan telah ditambah dalam bahan api, seperti perencat kakisan, pencuci, dan penambahbaikan pelinciran, untuk meningkatkan prestasi enjin dan perlindungan komponen enjin. Diesel bercampuran dan tiada campuran metil ester dari minyak sawit juga ditentukan secara kualitatif menggunakan PCA berdasarkan kehadiran metil ester asid lemak yang signifikan di dalam diesel. Kajian ini mendemonstrasi strategi penentukuran pelbagai variasi untuk penentuan spektrometri infra-merah serentak dari sifat fizikal dan kimia produk petroleum, iaitu, suhu mendidih pada takat 95% perolehan semula (T95%), takat kilat (FP), titik awan (CP) dan indeks setana (CI) yang merangkumi pemilihan rantau spektrum, penentukuran/pengesahan set pecahan, pra-pemprosesan data, dan regresi. Berdasarkan hasil ujikaji, penentukuran yang dibina di rantau gabungan 4800-4000 cm^{-1} menggunakan set penentukuran yang dipilih secara rawak, berjaya menyampaikan prestasi ramalan yang sangat baik dari segi pekali penentuan, punca min kesilapan persegi silang, punca min persegi kesalahan ramalan dan nisbah sisihan prestasi. Selain itu, kesemua model yang dihasilkan memenuhi keperluan reproduktif kaedah standard ASTM tanpa mengira pengambilan pembetulan penyebaran berbilang/*Savitzky-Golay* derivatisasi kedua dan regresi komponen utama/regresi. Ini menunjukkan bahawa kesesuaian model bergantung kepada setiap komponen penentukuran tunggal. Hasil kajian ini juga mendapati bahawa pra-rawatan data penting dalam menghasilkan ramalan-pelaksanaan ramalan.

Kata kunci: kualitatif, kuantitatif, kimometrik, produk petrolium

ACKNOWLEDGEMENTS

First and foremost, many thanks to ALLAH for giving me the strength, patience, and energy to complete my MSc research program. This thesis marks the culmination of a series of rather extraordinary paths in my life besides the uncertainty of my health condition and a tight schedule with daily work commitments. Therefore, it would not be complete without acknowledging my appreciation to all those who have helped me grow academically, professionally, and personally.

I would like to express my sincere gratitude to my supervisors, Professor Dr Sharifuddin Md Zain and Associate Professor Dr Low Kah Hin, for the continuous support of my MSc study and research for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research work and writing of the thesis. I could not have imagined having better supervisors and mentors for my MSc study.

Besides my supervisors, I would like to thank my beloved colleagues for their constant and continuous support throughout the research project, besides my busy schedule with my daily work. They provided me with very kind assistance in collecting samples and obtaining the FT-NIR and laboratory reference method analysis. Their full support and assistance made my work much more manageable and completed per agreed timelines.

Last but not least, I would like to thank my beloved family, especially my younger sister Fatimah Othman supporting me spiritually throughout my research project. Also, to my niece Azreen Rohani for her assistance in preparing the presentation pack during the international conference and candidature defense sessions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AI	:	Artificial Intelligence
ANN	:	Artificial Neural Networks
API	:	American Petroleum Institute
ASTM	:	American Society for Testing and Materials
CCI	:	Calculated Cetane Index
CDU	:	Crude Distillation Unit
CFPP	:	Cold filter plugging point
CI	:	Cetane index
CN	:	Cetane Number
CP	:	Cloud point
DA	:	Discrimination Analysis
FAME	:	Fatty acid methyl ester
FBP	:	Final boiling point
FCC	:	Fluid Catalyst Cracker
FP	:	Flash point
FTIR	:	Fourier Transform Infra-Red
FTNIR	:	Fourier Transform Near Infra-Red
HCA	:	Hierarchical Cluster Analysis
HDPE	:	High Density Poly- Ethylene
InAs	:	Indium Arsenide
LDA	:	Linear Discriminant Analysis
LWR	:	Locally Weighted Regression
LV	:	Latent Variable
MAX	:	Maximum

MID-IR	:	Mid Infra-Red
MIN	:	Minimum
MLR	:	Multiple Linear Regression
MS	:	Malaysia Standard
MSC	:	Multiplicative Scatter Correction
MToe		Million tonnes of oil equivalent
MW	:	Megawatt
NIPALS	:	Non-Linear Iterative Partial Least Squares
NIR	:	Near Infra-red
PC	:	Principal Component
PCA	:	Principal Component Analysis
PCR	:	Principal Component Regression
PLS	:	Partial Least Square
PLSR	:	Partial Least Square Regression
PMCC	:	Pensky-Martens Closed Cup
PME	:	Palm Methyl Ester
ppm (wt)	:	Parts per million in weight
RMSECV	:	Root Mean Square Standard Error Cross Calibration
RMSEP	:	Root Mean Square Error Prediction
RPD	:	Residual Predictive Deviation
R^2	:	Regression Coefficient
SEC	:	Standard Error of Calibration
SEL	:	Standard Error of Laboratory
SECV	:	Standard Error of Cross-Validation
SIMCA	:	Soft Independent Modeling of Class Analogy
SGSD	:	Savitzky Golay Second Derivative

SNV	:	Standard Normalized Variate
SD DEV	:	Standard Deviation
SVMR	:	Support Vector Machine Regression
T95	:	Boiling point at 95% recovery
ULSD	:	Ultralow-sulfur diesel
vol/vol%	:	Volume per volume percentage

CHAPTER 1: INTRODUCTION

The development of an economy is significantly supported and accelerated by energy. Since 2010, Malaysia's overall energy usage has been rising annually, with an exceptional reduction in 2015. The demand for energy in Malaysia is anticipated to increase annually, even as energy consumption for the upcoming years continues to rise. Suruhanjaya Tenaga (2019) estimates that Malaysia's energy consumption has grown at a pace of 50.7% over the eight years from 2010 to 2017 (Dzulkefli & Saad, 2020).

The three types of primary energy production in Malaysia are coal, natural gas and petroleum. Figure 1.1 shows the production of energy sources in Malaysia from 1980 until 2009. In 2004, petroleum reached its peak production with 861.8 thousand barrels daily however slowly declining in the following years. Contradicting to natural gas which inclines in production (Shafie et al., 2011). It is anticipated that oil imports would begin in 2013 and is expected to reach a total of 45 Mtoe by 2030 (Gan & Li, 2008).

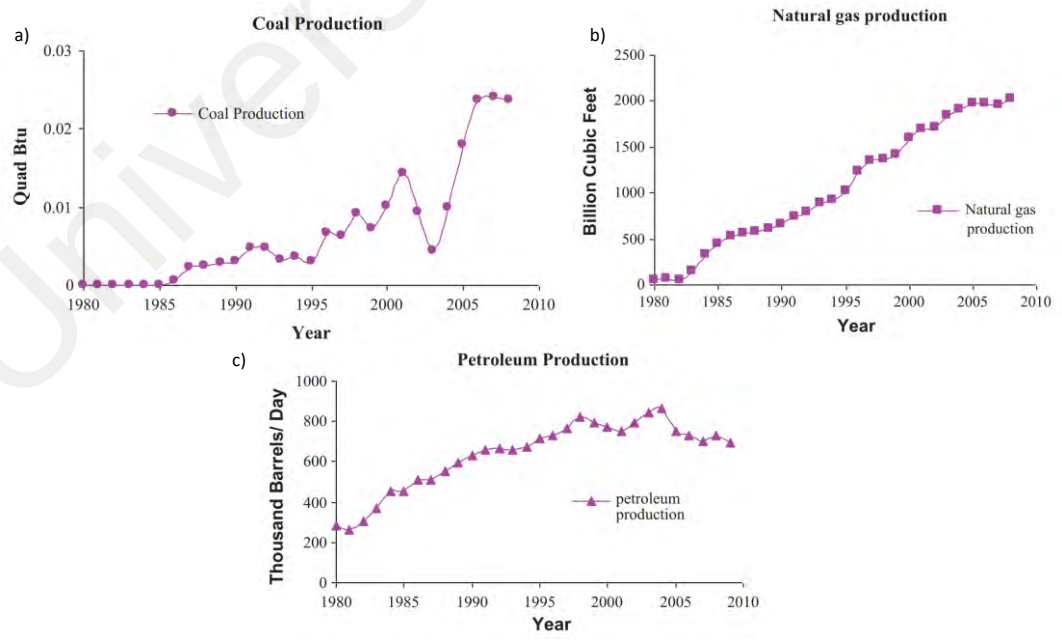


Figure 1.1: Production of primary energy sources in Malaysia; a) Coal production, b) Natural gas production and c) Petroleum production (Shafie et al., 2011)

Energy demand in Malaysia on 2009 since 1999 increases at about 66.5% from 9690 MW to 16,132 MW. The fast rate of economic development in Malaysia is the cause of this sudden rise in demand. Malaysia has a population of 25.4 million people as of 2009, and by 2020, about 75 percent of the country's population would reside in urban areas, more than doubling since 1980. This shows the significance of energy sources in Malaysia for both production and consumption as seen in Figure 1.2.

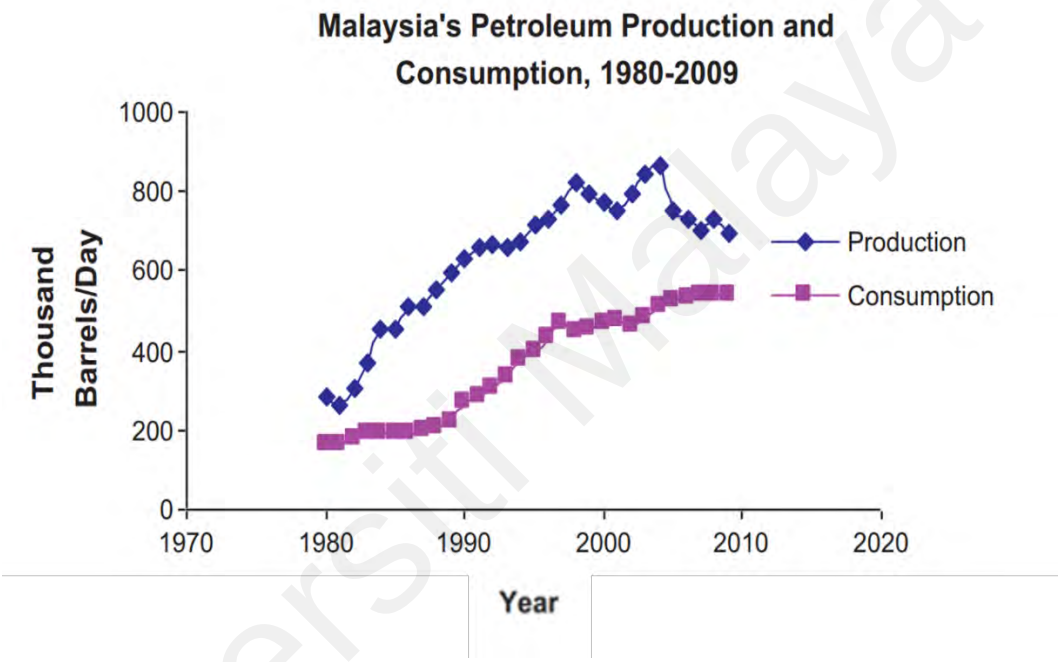


Figure 1.2: Malaysia's Petroleum Production and Consumption (Shafie et al., 2011)

Malaysia is popular for its abundance of oil reservoirs with one of the leading companies in oil and gas being Petroliam Nasional Berhad (Petronas). Based on Chong et al.'s (2015) study, six refineries in Malaysia where three being from Petronas has a total estimated volume processing approximately 500,000 bbl/d of crude oil. The most significant oil products in Malaysia's energy consumption, particularly in the transportation sector, were gasoline and diesel as seen in Figure 1.3.

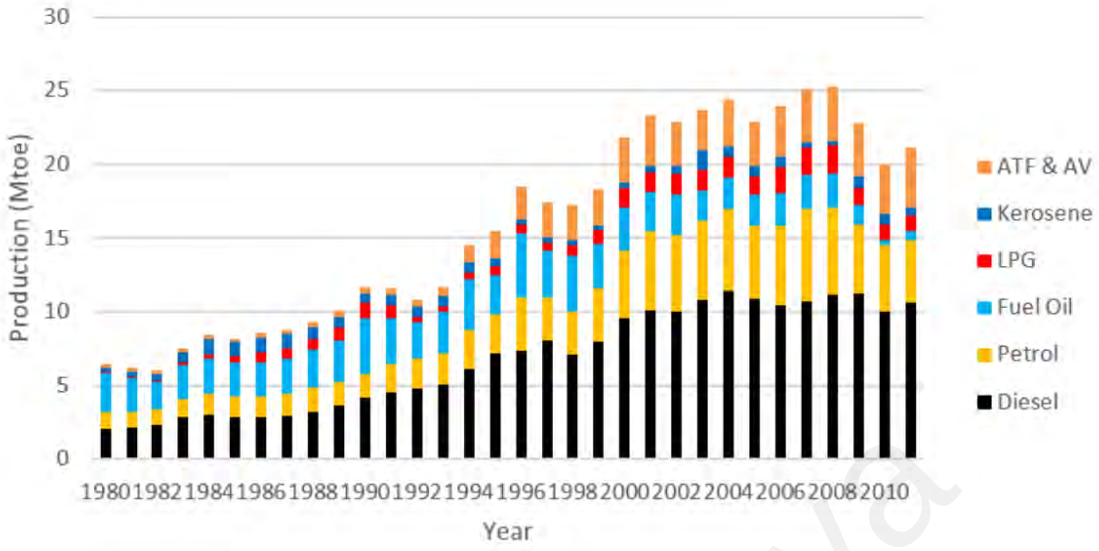


Figure 1.3: Oil production in Malaysia (Adapted from Chong et al., 2015)

1.1 Oil Refining

Refining is the process of separating crude oil into various ranges and hydrocarbon products based on their differences in boiling point. The processing needs the utilisation, for instance of heat exchangers, variation of pressure, temperature, and injection of chemicals where the large molecules from crude oil are separated into groups of similar molecules. The process additionally organizes their configurations and integration into diverse hydrocarbon substances and compounds into three major hydrocarbon groups: paraffinic, naphthenic, and aromatic (Altgelt, 1993).

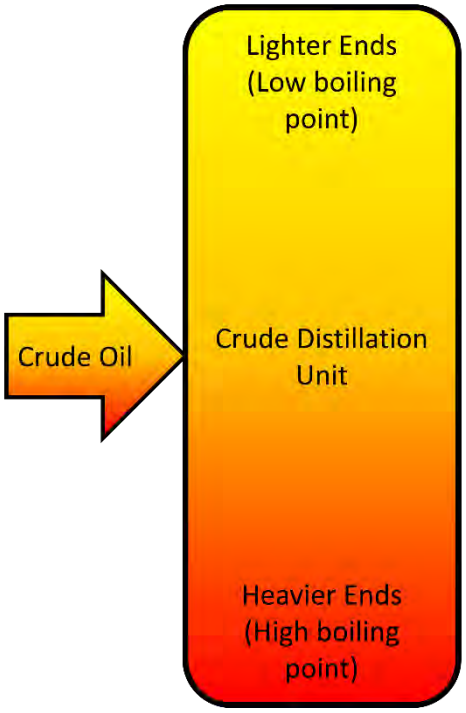
Refineries produce various products such as jet fuel, gasoline and diesel, including many required feedstocks for the petrochemical industry (Gary et al., 2007). Complex and integrated refineries incorporate treatment, fractionation, conversion and blending operations, including petrochemical processing. Further conversion processes convert the light distillates into more value-added products by changing the hydrocarbon molecules' structure and size via cracking, reforming, and other conversion processes. A typical

refinery process capacity is in the range of 100,00 barrels daily to 400,000 barrels daily depending on the capacity and complexity of the processing unit (Verma et al., 2017). Various separation methods, such as hydrotreating, extraction and sweetening, are used on different process streams to eliminate unwanted components and enhance the quality of the end product. In general, petroleum refining operations can be grouped into fractionation (distillation), light gas oil processing, heavy oil processing, and treatment and environmental protection processes (Rana et al., 2007).

Prior to the crude oils being fed at the Crude Distillation Unit (CDU) tower, the crude treatment, i.e., desalting process using electrostatic or chemical separation, is essential to remove undesirable constituents such as inorganic water salts, water, water-soluble trace metal contaminants and suspended solids.

As illustrated in Figure 1.4, the desalted crude is processed at pressures slightly above atmospheric and temperatures ranging from 345 to 370 ºC via distillation column tower. To avoid thermal cracking, the residue obtained from atmospheric distillation will be processed via a vacuum distillation tower at reduced pressure and elevated temperatures. (Aitani, 2004).

Crude oil distillation onto straight-run cuts occurs in both vacuum and atmospheric towers. The different types of crudes have various characteristics directly related to the composition of the products and their production yield. The main fractions obtained have specific boiling point ranges and can be categorized based on the decreasing of boiling point and increasing of the molecular weight into light distillates (naphtha), gases, middle distillates (kerosene and diesel/gas oils) and heavy distillates (heavy gas oils and residue) (Mochida et al., 2014).



Boiling Range		Cut side products	
→	< 29 °C	→	Butane and lighter
→	29 – 85 °C	→	Gasoline blending
→	85 – 176 °C	→	Naptha
→	176 – 232 °C	→	Kerosene, jet fuel
→	232 – 343 °C	→	Distillates (diesel, heating oil)
→	343 – 565 °C	→	Heavy gas oil
→	>565 °C	→	Residual fuel oil

Figure 1.4: Crude distillation unit and cut points for each intermediate product (Adapted from Balaji, 2017).

1.2 Crude Oil Products

The elemental composition of crude oil depends on the source and type of the crude; nonetheless, these elements exhibit slight variations within a narrow range. (Riazi, 2005). A specific crude oil is not easily identifiable or a quantifiable compound. Paraffin, aromatics and napthenes plus a tiny number of organic compounds such as oxygen, sulfur nitrogen, and traces of metallic compounds such as nickel, vanadium and sodium, are typical crude oil mixtures of hydrocarbon compounds. Typically, these compounds with less than 16 carbon numbers are made up of a relatively high proportion of crude oil, comprising 84.5% carbon, 1-3% sulfur, 13% hydrogen and less than 1% each of oxygen, nitrogen and metal salts (Aitani, 2004).

Crude oils can be categorised in many ways, either by their physical or chemical properties, commonly based on their density, American Petroleum Institute (API), sulfur

content, and hydrocarbon composition, typically in the range of 50° API to 10° API (Ilyin et al., 2016). In terms of its physical properties, crude oil can be distinguished by API gravity; the greater the API gravity, the lighter the crude. Crude oils with high hydrogen, low carbon and high API gravity are usually rich in paraffin and tend to yield more significant proportion of light petroleum products and gasoline. In contrast, those with low hydrogen, high carbon and low API gravity are usually rich in aromatics (Demirbas et al., 2015). Crude oils with a low sulfur content of less than 0.5 wt% and high sulfur content of more than 0.5 wt% are defined as sweet crude and sour crude, respectively (Ghulam et al., 2013).

1.3 Diesel Fuel

The primary oil products consist mainly of transportation fuels, which account for around 52% of global oil usage. Hence, the sustainability and availability of gasoline and diesel petroleum-based products as transportation fuels are the main concerns in the global market for the future.

Besides gasoline, many petroleum products derived from is diesel. For the non-complex refinery, diesel is derived as a single component (straight run product). However, for complex refineries, it is derived from blending operations. The diesel blended source is typically obtained from atmospheric distillation, hydrocracking, distillate hydro-treater, Fluid Catalyst Cracker (FCC) light cycle oil, and several products from vacuum distillation and delayed coker unit (Jones, 2008). Different crude oil will produce different quality of diesel rundown. The challenges are to blend different quality of diesel rundown, to meet final product diesel specification which require rapid online assessment. (time consuming, high reblend rate, high product giveaway, high demurrage cost).

The sustainability of fuels globally at present is a concern. Most oil companies and refineries have started blending the crudes with bio-renewable resources such as algae,

recycled waste, and palm oil. Similar to other petroleum products, diesel and jet fuels are also blended with bio-based components with an acceptable percentage without jeopardising the final products' quality of meeting the fuel specifications and engine performance. In Malaysia, biodiesel was gazetted by the government as being of five per cent fatty acid methyl ester (derived from palm oil) blended with diesel petroleum.

The key characteristic of diesel fuel used in automotive engine combustion is its cetane number, which indicates the ease of engine ignition and combustion (Barabas et. al., 2010). The essential properties of diesel fuels in process control and meeting product specifications are boiling point at flash point, 95% recovery, cetane index, cloud point and total sulfur. Diesel fuel and domestic heating oil have an approximately 200 to 375^{o} C boiling point range. Total sulphur reduction and cetane improvement are required for ultralow-sulphur diesel (ULSD) product grade production. Substantial investment in hydrotreating will be necessary to meet upcoming ULSD specifications, which demand sulfur content is between 10-15 ppm. (Stuntz & Plantenga, 2002).

1.3.1 Diesel compositions

Diesel fuels are intricate mixture of hydrocarbon molecules, generally boiling within the temperature range from 150^{o} C to 380^{o} C. They are typically blended from several refinery streams, mainly from the primary distillation unit. However, components from other units often increase diesel fuel production in a conversion refinery. The proportion of cracked gas oil components in blended diesel is typically low since the high aromatic content of the cracked gas oil lowers the cetane value of the blended diesel fuel (Gary et al., 2007; Parkash, 2003).

Diesel fuel, with the carbon number ranging from about C8 to C24, comprises approximately saturated hydrocarbon derivatives (75% v/v, primarily paraffin

hydrocarbons including iso-paraffins, n-paraffins, and cycloparaffins), and aromatic hydrocarbon derivatives (25% v/v, including alkylbenzenes and naphthalene derivatives). Diesel fuels primarily contain a mixture of C_{10} to C_{19} (Retnam et al., 2015).

Diesel fuel's general appearance, or colour, is a valuable indicator to identify contamination by residual (higher boiling point) constituents, fine solid particles or water. Therefore, it is prudent to check that visual inspection delivers clear fuel. Since the color of diesel fuels is utilized for manufacturing control objectives, it is crucial to determine it as part of the fuel's appearance. Typically, the methods require a visual determination of colour using coloured glass discs or reference materials. The colour may indicate the degree of refinement of the material. Similarly, odour is vital when it comes to acceptance. It is usually required that diesel fuel is reasonably free of contaminations, such as mercaptan derivatives (RSH, also called thiol derivatives), which impart unpleasant odours to the fuel.

Table 1.1 shows the Malaysia Standard Diesel fuel specification for Euro 5 which shows the maximum boiling point at 95%, minimum flash point, maximum cloud point and minimum cetane index.

Table 1.1: Malaysia Standard High PME Diesel Fuel Specification Euro 5

Properties	Minimum	Maximum	Referee Test Method
Boiling point at 95% recovered volume, °C	-	360	(ASTM D86, 2020)
Flash point, °C	60	-	(ASTM D93, 2020)
Cloud point, °C	-	19	(ASTM D2500, 2017) / ASTM D 5772
Cetane index	49	-	(ASTM D976, 2016) / ASTM D4737

1.3.1.1 Volatility / Boiling point

The volatility properties of a diesel fuel can be described by its boiling temperature. A standardized device is used to heat a fuel sample and distill successive fractions of diesel fuel under controlled conditions.

The most standard test method used for the distillation is ASTM D86 (Leonardo et al., 2020). During the distillation analysis, initial boiling point (IBP), endpoint (EP) or final boiling point (FBP), per cent of condensate recovered and per cent residue of non-volatile matter data information will be recorded (Zvirin et al., 1998).

The diesel fuel's volatility or boiling range impacts many other properties, including flash point, density, viscosity and cold properties (cloud point, pour point) (Bahadur et al., 1995), which must meet the diesel fuel specification. High volatility or low boiling point could cause vapour lock and lower flashpoint, which are not desirable qualities. Vapour lock can cause engine performance to deteriorate, where misfiring or failure to restart can occur after a brief shutdown in hot conditions (Thomas, 1988; Chang et al., 2020). However, in the combustion chamber, greater volatility allows for more complete vaporization of the fuel.

As a result, components with high boiling points may not combust fully, resulting in the formation of engine deposits and elevated levels of certain substances. Within the range of 350ºC to 400ºC, however, the effect on the exhaust emission is considered relatively low at low volatility or high boiling. The effect on the tendency to smoke via influence on the injection and mixing the fuel at mid-volatility, boiling point at 50% recovery of diesel fuel also has a marked effect. There is also an interest in the 50% distillate recovery temperature for the calculation of the cetane index (CI) by ASTM D976 (2020).

It is emphasised that in practice, the mix of volatilities is very important. High volatility components at a lower boiling point improve cold starting and warm-up, while low-volatility components at a higher boiling point tend to increase wear, smoke and deposits.

ASTM D975-2020 is the international standard method specified in the diesel fuels specification. The most critical volatility of diesel fuel specification is boiling point at 95 % recovery (T95). For Malaysia Standard High PME Diesel Fuel Specification Euro 5 (MS 123-5, 2020), the T95 is 360 ^{o}C maximum. A typical diesel fuel distillation curve is illustrated in Figure 1.5.

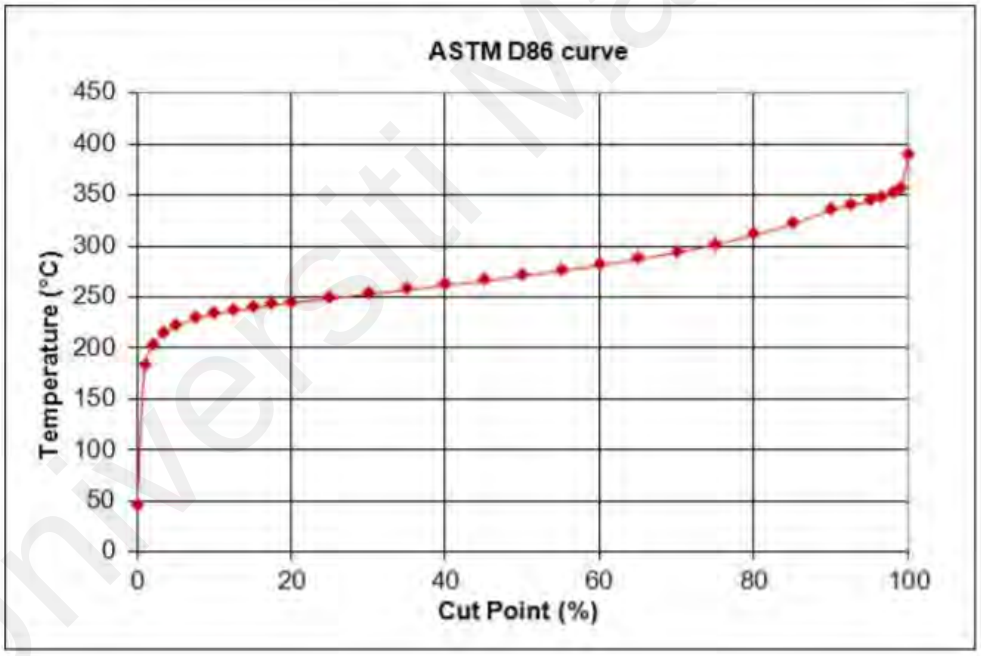


Figure 1.5: Distillation curve for diesel fuels. (Adapted from Santos et al., 2021)

1.3.1.2 Flash point

Flash point is one of the vital properties in diesel fuels that concerns safety while handling, transporting and storing. The flash point test will measure the lowest temperature when the diesel fuel produces enough vapour to cause ignition leading to

flame generation. The flash point average for diesel fuel ranges from 55°C to 66°C. Fossil diesel is composed of low molecular weight molecules and a branched compound that reduces the flashpoint (Knothe, 2010).

The standard test method ASTM D93 (2020) for flash points by Pensky-Martens Closed Cup Tester for diesel fuels is commonly used to determine flash point quality. ASTM D975-2020 is the international standard method specified in the diesel fuels specification. The flash point specification is in the range of a minimum of 30ºC to 55ºC depending on the grades of diesel fuels. For Malaysia Standard High PME Diesel Fuel Specification Euro 5 (MS 123-5, 2020), the flashpoint specification is 60.0 ºC at the minimum limit.

Flash point provides an excellent indication of diesel fuel contamination with more volatile products such as contaminated with gasoline or fuel oil and adulteration of diesel blended with other blending components illegally. The significance of the flash point test is for the safety requirements during storing and handling. (Kaisan et al., 2017).

1.3.1.3 Cloud point

Cloud point is the temperature at which a surfactant solution starts to molecularly agglomerate generating a cloudy physical appearance (Swarup and Schoff, 1993). About 20% of the diesel fuel composition causes solubility limitation due to the relatively high heavy paraffinic hydrocarbons (Mohammadi et al., 2022). The paraffins are most likely to deposit out as wax in adequate cool condition. The wax formation is undesirable because they contribute to high cetane numbers. The wax formation in a vehicle fuel system is a potential source of operation problems (Dwivedi & Sharma, 2014). Wax-related tests are used to determine the low-temperature properties of fuel, including the pour point (PP), cloud point (CP) and cold filter plugging point (CFPP) (Das et al., 2022). The standard method used for CP analysis is ASTM D2500 (2020) which is the test

method that describes the temperature at which the cloud becomes visible when the fuel is cooled.

ASTM D975-2020 is the international standard method specified in the diesel fuels specification. However, there is no specific or standardized CP specification for all grades of diesel fuels. It is unrealistic to specify low-temperature properties to ensure satisfactory operation under all ambient conditions, especially in cold countries. For Malaysia Standard High PME Diesel Fuel Specification Euro 5 (MS 123-5, 2020) specified, the CP specification limit is a maximum of 19.0 ^{o}C.

1.3.1.4 Cetane number/ Cetane index

The accepted international standard method used to measure the ignition quality of diesel fuel is the Cetane Number (CN), which measures the delay of ignition of the diesel fuel. It is based upon the ignition characteristics of two hydrocarbons, 2,3,4,5,6,7,8-heptamethylnonane and n-hexadecane (Cetane). It measures the ignition quality of diesel fuels and the reference to the blended fuel percentage of cetane blended with heptamethylnonane, which matches the ignition quality of the test fuel (ASTM D613, 2019). A cetane number of 100 is assigned to cetane, which has a brief ignition delay period. On the other hand, heptamethylnonane has a long delay period and is given a cetane number of 15. The engine used in cetane number determination is a standardised single-cylinder, variable compression ratio. The engine, the operating condition, and the test procedure are specified by ASTM D 613 test method (Barabas et al., 2010).

Due to the high cost of the cetane number determination, an alternative method often used for cetane determination is based on the CI formula. This method estimates the cetane number of diesel fuels from API gravity and mid-boiling point. As computed from

the formula, the index value is designated as a calculated cetane index (ASTM D976, 2020). The CI usually gives values slightly above the CN. However, it provides a good indication of ignition quality in many cases but the CI will be less accurate when the cetane improver additive is added to diesel fuels (ASTM D976, 2020). A high cetane number indicates that the fuel ignites more readily when sprayed into hot compressed air.

ASTM D975-2020 is the international standard method specified in the diesel fuels specification. The CN and CI range of 40 to 47 is usual for all grades of diesel fuels. For Malaysia Standard High PME Diesel Fuel Specification Euro 5 (MS 123-5, 2020) specified, the CN and CI specification limit is a minimum of 49.0.

The significance of the cetane number or calculated cetane index is that they indicate engine fuel performance. Increasing the cetane number of diesel fuel can result in improved cold-starting performance, reduced smoke emission during warm-up, reduced noise, lower fuel consumption and reduce exhaust emissions. As a result, certain countries strive to raise the cetane number of their diesel fuel.

1.4 FT-NIR Spectroscopy Applications in Refineries

The current practice in the oil industry, i.e., in refinery's quality assurance and quality control laboratories relies on the use of conventional and univariate method measurements in the quantitative and qualitative assessment of their feedstock, process samples, intermediate products, and final products blending control and optimisation.

This conventional quality measurement methodology consumes a high workforce, longer analysis time, high operating expenditure, and capital expenditure cost, leading to low productivity, low operational efficiency, and low profitability.

In refinery processing plants, fast and immediate process control and optimisation are critical in achieving high production volume and high-quality products that meet sales specifications and increase profitability, productivity, and operational excellence.

A new innovative way of qualitative and quantitative measurement and application needs to be explored to improve and address the pain points and provide a total solution aligned within current trends and practices, i.e., the application of chemical analytical strategies with data analytics is imperative.

The new method involves using data analytic methodologies, i.e., Chemometrics Multivariate methodology, and a spectroscopic technology, i.e., Fourier Transform Near Infra-Red (FT-NIR) Spectroscopy. This transforms the conventional method of process control and petroleum product blending derived from complex multiple blending components with various compositions, ratios, and formulations to predict the qualitative and quantitative product quality, including chemical and physical properties.

The physical and chemical data derived from the various reference analytical methods and FT-NIR spectra will then be analysed using various data analytic techniques to determine the quality of the blended petroleum products. The capability of FT-NIR for fast, non-destructive, simultaneous multi-compositional and online analyses allows for more repeatable measurements and faster compared to conventional laboratory and online process analysers. It also enables the measurement of several properties simultaneously with only one FT-NIR analyser on an almost real-time basis (less than 1 minute).

This research study describes the use of FT-NIR technology with various chemometric methods in petroleum products i.e., gasoline, diesel, and kerosene. This study covers the FT-NIR spectra pre-treatment Multiplicative Scatter Correction (MSC) and Savitzky Golay Second Derivative (SGSD), qualitative Principal Component Analysis - Soft

Independent Modeling of Class Analogy (PCA - SIMCA), and quantitative regression techniques namely Principal Component Regression (PCR) and Partial Least Square Regression (PLSR) for petroleum products (gasoline, kerosene, and diesel).

The statistical diagnostic tools, i.e., Root Mean Square Standard Error Cross Calibration (RMSECV) and Root Mean Square Error Prediction (RMSEP) are used for calibration and validation model performance. A comparison with the standard univariate conventional reference method reproducibility is also made.

For refiners, the FT-NIR technology is the most selective spectroscopy technology deployed for qualitative and quantitative measurement of petroleum hydrocarbon application. The FT-NIR wavelength region is located between the electromagnetic spectrum's visible and mid-infrared (MID-IR) regions, ranging from 800 to 2500 nm (12,500 to 4000 cm^{-1}). NIR absorption bands region is related to the overtone and combination bands of the fundamental vibrations of -CH, -NH, -SH, and -OH groups in the MID-IR (Start et al., 1986; Weyer, 1985; Bunasiu et al., 2015)

NIR region analysis is especially advantageous in situations where chemical analysis is limited by light penetration. Compared to MID-IR, NIR radiation can penetrate deeper. Generally, the entire NIR region exhibits clear spectral differences among paraffinic, naphthenic, and aromatic hydrocarbons. However, the 5000-4000 cm^{-1} (combination bands region) provides the most informative spectral features regardless of the colour of the hydrocarbon samples.

The selection of the overtone region is vital for the chemometrics multivariate calibration model's development. This is to ensure excellent signal-to-noise ratio spectra, higher quality NIR spectra with good reproducibility and eliminate loss of information from the spectra data. From the literature review, the combination overtone region in the

range of wavenumber 4800 to 4000 cm^{-1} with the cell path length 0.5 mm is the most suitable to cover any refineries' petroleum hydrocarbon physical appearance ranging from clear liquid to dark in colour. With this combination, the overtone region will eliminate interference such as excessive water presence and fluorescence or scattering effect due to coloured and dark petroleum products.

Refineries' petroleum hydrocarbon is a highly complex mixture of hydrocarbon and heteroatomic organic compounds of varying molecular weights and polarities. The refineries' petroleum products' physical appearance range from clear liquid to dark in colour, corresponding to the increment of carbon atoms (molecular weight) and complexity of the composition. This research study covers four properties of diesel for quantitative measurement, which are boiling point at flash point (FP), 95% recovery (T95), cetane index (CI) and the cloud point (CP).

1.5 Chemometrics

Chemometrics is a chemical field that employs mathematical, statistical, and other methodologies to apply formal logic in designing or selecting optimal measurement procedures and experiments, as well as to extract the most pertinent chemical information by analyzing chemical data (Burgard & Kuznicki, 2018).

It is an application-driven discipline that addresses both predictive and descriptive problems related to chemical data. Descriptive analyses involve the examination of the latent structures present in the data, whereas predictive analyses model the data to estimate the desired properties of the targeted system.

The food and agriculture industries have been among the earliest application areas and driving forces behind the development of chemometrics. Many of the earliest

chemometricians were from food and agricultural research institutes (Brereton, 2003). The early applications mostly involved multivariate classifications. However, the applications have diversified substantially over the years. It was due to research interests and partly to economic driving forces. Most of their applications reported in the academic literature are often related to experimental design and multivariate data interpretation (Loh, 2016).

Several spectroscopic and chromatographic methods can provide analytical data on many components of a single specimen. Multivariate data needs several variables to be measured for each specimen. One of such example data in analytical chemistry is discrimination; such as the investigation of an oil spill to determine the particular source of the pollution by analysing the fluorescence spectrum (Miller & Miller, 2018). While it is feasible to compare specimens by examining each variable, modern computers enable more advanced processing techniques where all variables can be simultaneously taken into account.

Variables can be divided into two groups: predictor variables and response variables. The situation in which we have a response variable y, depending on several predictor variables, x_1, x_2, x_3,…., can be studied as a multiple regression. A simple example would be when y is an absorbance value from the mixture of compounds with concentrations x_1, x_2, x_3,… The technique of linear regression can be extended to find a regression equation in the form:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots b_nx_n \quad (1.1)$$

The prediction performance can be validated in a way similar to the validation of linear discriminant analysis (LDA), i.e. either by dividing the data into two randomly chosen groups, making the model with one group and then testing it with the other or by using a

'leave-one-out' method. A graph of the predicted values against the measured values gives a point close to a straight line if the model is satisfactory (Miller & Miller, 2018).

1.5.1 Qualitative multivariate measurement

Qualitative data refers to data which could be observed and not measured. Multivariate analysis is a complex from of statistical analytical technique used when there are more than two variables in the data set (Mertler et al., 2021). In this section, one type of qualitative multivariate measurements will be explained and explored which is the principal component analysis.

1.5.1.1 Principal component analysis (PCA)

The challenge with multivariate data is that the large amount of data may make it challenging to identify patterns and relationships. For instance, a spectrum is typically described by numerous intensity or absorbance measurements, rather than just a few variables, resulting in a correlation matrix that contains hundreds of values. Consequently, many multivariate analysis methods aim for data reduction.

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) reduces the dimensionality of such datasets, increasing interpretability while minimising information loss. It creates new uncorrelated variables that successively maximise variance (Jolliffe & Cadima, 2016).

The idea behind PCA is to find the principal components $Z_1, Z_2, \dots, Z_n$, which are linear combinations of the original variables describing each specimen, $X_1, X_2, \dots, X_n$, i.e.:

$$Z_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1n}X_n \quad (1.2)$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{2n}X_n \quad (1.3)$$

and so on. The coefficients a_{11}, a_{12}, …, are chosen so that the new variables are not correlated, unlike the original variables. The principal components are also chosen. The first principal component (1 PC), Z_1, explains most of the variation in the data set, the second (2 PCs), Z_2, explains for the next largest variation and so on. Hence, useful PCs are much less than the number of original variables (Miller & Miller, 2018).

A different approach is needed to avoid an unknown object being incorrectly classed. It is possible that the unknown object may not fit into any of the classes being considered. Hence, it requires a rule which allows discrimination between membership and non-membership of a given class. This problem can be solved by making a separate model for each class and using the model to test whether the unknown object could be a class member. As such, the Soft Independent Modeling of Class Analogy (SIMCA) method makes a model of each class in terms of the first few principal components for data reduction with more variables like spectroscopic data (Miller & Miller, 2018).

1.5.2 Quantitative multivariate measurement

The term "quantitative data" refers to data that may be measured numerically with the intention of doing statistical analysis on the resulting data. This section covers a few quantitative multivariate measurements used in this study which are partial least squares regression (PLSR) and principal component regression (PCR).

1.5.2.1 Principal component regression (PCR)

One problem in multiple regression is that correlation between the predictor variables can lead to mathematical complications, resulting in an unreliable prediction of y (Slinker & Glantz, 1985). A solution is to carry out a PCA on the x variables and then regress y

on the principal component. Since the principal components are not correlated, the correlation between the predictor variable is overcome. This method is known as PCR.

PCR is the reduction of predictor variables by using the first few PCs rather than the original variables. This method will give satisfactory results provided that the PCs used account for most of the variation in the predictor variables. PCR is also a valuable technique when the number of original predictor variables exceeds the number of calibration specimens available (Miller & Miller, 2018).

1.5.2.2 Partial least squares regression (PLSR)

PLSR starts by finding a linear combination of the predictor variables. However, how these linear combinations are chosen is different from PCR. In the PCR, PCs are chosen to explain the largest variation in predictors without considering the intensity of the relationships between the predictor and response variables. However, in PLS, extra weight is given to the predictor variables that correlate with the response variables; hence the prediction is more effective.

By doing this, the linear combinations of the predictor variables that are strongly correlated with the response variables and account for the variability in the predictor variables are selected. A distinction is usually made between the situation when the response consists of a single variable and when the response is multivariate.(Miller & Miller, 2018)

1.5.3 Advantages of multivariate analysis combined with FT-NIR spectroscopy applications for the refinery

The implementation of FT-NIR spectroscopy in petroleum industries has dramatically increased over the past 15 years (Chung, 2007). FT-NIR has fast, non-destructive, online, and simultaneous multi-compositional analyses. It grants for a faster and more repeatable

measurements than conventional offline and online analysers, such as flash point analysers. It also enables the measurement of several properties simultaneously using only one NIR analyser on an almost real-time basis of less than 1 minute. Online measurements using NIR are particularly relevant since the improved harmonisation of real-time analysis and process control can gain significant economic benefits (Chung, 2007)

1.6 Research Objective

This research proposes to conduct; Firstly, qualitative measurement, to differentiate a) petroleum products between gasoline, diesel and kerosene, b) gasoline without additive, and gasoline with additive. c) diesel without palm methyl ester (PME) and diesel blended with PME. Secondly, quantitative measurement, to measure the physical and chemical properties of the diesel petroleum product obtained from various local refinery plants. The analytical reference methods include physical distillation, flash point, cloud point, cetane index, and FT-NIR analyser. The physical and chemical data derived from the various reference analytical methods and FT-NIR spectra were analysed using multiple data analytic techniques to determine the quality of the diesel petroleum products. The FT-NIR spectra are essential to assess the quality of petroleum products. The detailed objectives of this research are as follows:

1. To perform FT-NIR spectra pre-treatment or pre-processing using different pre-treatment algorithms, such as Multiplicative Scatter Correction (MSC) and Savitksky-Golay Second Derivative (SGSD).
2. To develop a qualitative model to discriminate between the following based on the FT-NIR measurement using Principal Component Analysis (PCA).
    a. Gasoline, kerosene and diesel
    b. Gasoline with and without additives

c. Diesel with and without palm methyl ester (PME)

3. To develop quantitative calibration models for the boiling point at 95% recovery (T95), Flash Point (FP), Cloud Point (CP), and Cetane Index (CI) using different FT-NIR spectra pre-treatment algorithm types, i.e., MSC and SGSD and multivariate regression methods, i.e., PCR and PLSR.
4. To evaluate the accuracy and precision of quantitative calibration models' PCR and PLSR performance using MSC and SGSD in comparison with conventional univariate laboratory test methods.

CHAPTER 2: LITERATURE REVIEW

2.1 Application of FT-NIR spectroscopy in refineries.

Current practices in the oil industry, i.e., refinery quality assurance and quality control laboratory, are using conventional and univariate method measurements in the quantitative and qualitative assessment of their feedstock, process samples, intermediate products and final products blending control and optimisation.

This conventional quality measurement methodology consumes a large number of workforce, longer analysis time, high operating expenditure and capital expenditure cost, leading to low productivity, operational efficiency and profitability.

In the refineries processing plant, fast and immediate process control and optimisation are critical in achieving high production volume and high-quality products that meet sales specifications and increase profitability, productivity and operational excellence.

A new innovative way of qualitative and quantitative measurement and application needs to be explored to improve and address the pain points and provide a total solution aligned with current trends and practices, i.e., applying chemical analytical strategies with data analytics is imperative.

The new method involves using data analytic methodologies, i.e., Chemometrics Multivariate methodology and a spectroscopic technology, i.e., Fourier Transform Near Infra-Red (FT-NIR) Spectroscopy. This method transforms the conventional method of controlling process control and petroleum product blending derived from complex multiple blending components with various compositions, ratios and formulations to predict the qualitative and quantitative product quality, including chemical and physical properties.

The physical and chemical data derived from the various reference analytical methods and FT-NIR spectra will then be analysed using various data analytic techniques to determine the quality of blended petroleum products. The FT-NIR spectra are essential to assess the quality of the petroleum products either qualitatively or quantitatively, or both.

FT-NIR is a non-destructive and fast measurement tool which can read within less than a minute. It can measure several multi-parameters or properties (chemical or physical) simultaneously with acceptable accuracy and precision compared to conventional laboratory univariate and online process conventional analysers (Chung, 2007).

This research study describes the FT-NIR technology used in conjunction with several chemometrics methods covering FT-NIR spectra pre-treatment (MSC and SGSD), qualitative (PCA - SIMCA) and quantitative regression technique inclusive of PCR and PLSR for petroleum products (gasoline, kerosene, and diesel).

Root Mean Square Standard Error Cross Calibration (RMSECV) and Root Mean Square Error Prediction (RMSEP) are regularly used statistics. The calibration and validation model performance evaluation is based on statistical diagnostic tools. It was assessed comparatively with laboratory univariate conventional reference method for reproducibility.

2.2 FT-NIR technology

For refiners, the FT-NIR technology is the most selective spectroscopy technology that has been deployed for qualitative and quantitative measurement for petroleum hydrocarbon application. The FT-NIR wavelength region is located in the Infra-Red electromagnetic spectrum, covering the Near, Middle, and Far infra-red regions. The NIR absorptions consisted of overtone (first and second) and combination bands, corresponding to the vibrational of components functional groups -CH, Ar-H, -NH, -OH,

and -SH groups. The FT-NIR wavenumber is between 12,500 cm^{-1} to 4000 cm^{-1} (Stark et al., 1986; Weyer, 1985).

The NIR region (combination bands, first and second overtones) provides distinct spectral variations of hydrocarbon groups between paraffin, olefins, naphthene, and aromatics. The higher penetration depth for MID-IR as compared to NIR radiation, provides more advantages and benefits in analysis. However, the NIR combination bands region (4800-4000 cm^{-1}) provides the most information about spectral variations and features regardless of the colour of the hydrocarbon samples (Pasquini, 2018).

The selection of cell path length, cell measurement temperature, overtone bands region, and others are vital to acquiring excellent signal-to-noise ratio spectra, higher quality NIR spectra with good reproducibility, and eliminating loss of information from the spectra data.

The FT-NIR technology is a highly reliable spectroscopic analyser, which is widely used for various applications, specifically for refineries' petroleum hydrocarbons. FT-NIR spectroscopy can characterise and categorise the groups or types of petroleum products such as Gasoline, Diesel, Kerosene, and Fuel Oil. The fundamental measurement in obtaining information from the FT-NIR spectrum for the different types of chemical bonds of interest such as -CH, -C_2H_2, Ar-H, -SH, -OH, and others forms the basis in predicting petroleum products' quality properties.

Petroleum hydrocarbon composition is a highly complex mixture comprising of hydrocarbon and heteroatomic organic compounds that contributes to the variation of molecular weights and polarities. The refinery's petroleum products physical appearance ranging from clear liquid to dark in colour, corresponds to the increment of carbon chain which represents its molecular weight and the complexity of the composition. This

research study covers four properties of diesel for quantitative measurement, which are boiling point at 95% recovery (T95), flash point (FP), cloud point (CP), and cetane index (CI). FT-NIR are able to detect the chemical properties such as total sulphur, aromatics, benzene etc. and the physical properties of such component for example density, distillation and flash point. for refineries related application.

Selection of the overtone region is vital for the chemometrics multivariate calibration model's development to obtain the most accurate and precise measurement results. From the literature review, the combination overtone region in the range of wavenumber 4800 cm^{-1} to 4000 cm^{-1} with the cell path length 0.5 mm are the most suitable to cover any refineries petroleum hydrocarbon physical appearance ranging from the clear liquid until dark in colour (Pontes et al., 2011). In this combination, the overtone region will eliminate interference such as excessive water presence and fluorescence or scattering effect due to coloured and dark petroleum products.

2.3 FT-NIR spectra pre-treatment or pre-processing

From the research studies literature review, the application of chemometrics multivariate data analysis for many applications is based on the FT-NIR spectroscopy measurements in which acquired high-quality spectra data and signal-to-noise ratio.

During the FT-NIR spectra acquisition, there is an occasion it contains out-of-range values, very low or high absorbance values, missing absorbance values, and others (Ulmschneider & Roggo, 2008). These are probably caused by scattering or fluorescence effects, chemical interferences, or instrument drift (Broeke & Koster, 2019). Before analysing the data, FT-NIR spectra pre-treatment shall be performed to eliminate the complicated data analysis, and interpretation which will lead to misleading results.

The spectra pre-treatment reduces or eliminates the non-relevant spectral data information, which is vital to obtaining accurate and robust regression models. Hence, mathematical spectra pre-treatment shall be applied. There are various FT-NIR spectra pre-treatment techniques and methods typically used, such as Smoothing (Savitzky-Golay), first and second derivatives and normalisation such as Multiplicative Scatter Correction (MSC), Standard Normalized Variate (SNV), and others. On certain occasions, the spectra pre-treatment combination is also applied. However, two types of spectra pre-treatment were used for this research study: MSC and SGSD.

2.3.1 Savitzky-Golay – Smoothing and Derivatives

In 1964, Abraham Savitzky and Marcel J.E. Golay first described the Savitzky-Golay (SG) smoothing filter (Savitzky & Golay, 1964). By performing a local polynomial regression (of degree k) on a set of values (of at least k+1 points) that are used to smooth the data, the algorithm basically calculates the smoothed value for each point in the spectrum.

The window size and polynomial order selection are required for the algorithm application. If the polynomial order is low and the large window size, it contributes to high smoothing. SG method, in general, is averaging the data (including a subrange of data), which corresponds to using SG with zero polynomial order.

Smoothing is the simplest method to eliminate the signal-to-noise from the samples. The smoothing method considers the independent variables (FT-NIR spectra with 1.0 cm^{-1} interval) of the data matrix which contains similar information that can be averaged together, reducing the noise without significant loss of the signal of interest. The diesel spectrum uses 17 smoothing points to minimise and eliminate the signal noise for this research study.

In addition to the SG smoothing filter, most chemometricians will perform the FT-NIR spectra pre-treatment or pre-processing simultaneously with either the first or second derivative SG. This derivative method is commonly used to remove background signals and enhance the spectra features and visual resolution (Williams et al., 1992).

Baseline shifts and drifts of the numerical differentiation of digitised signals can be corrected depending on the order of the derivation. Derivative profiles are exhibited frequently with the increment of the resolution of the overlapping peaks and minor structural differences between nearly similar signals (Taavitsainen, 2009).

The first derivative is the most straightforward technique of a derivative. The spectral signal y = f(x) represents the response variable y, rate of change with independent variables x, where $y' = \frac{dy}{dx}$ represents the line tangent slope to the signal. Therefore, the first derivative, SG, can apply the baseline shifts correction. Further derivation, the second derivative, is given by $y'' = \frac{d^2y}{dx^2}$. There is a disadvantage in using derivative spectra pre-treatment, whereby the random noise will be increased, characterised by high-frequency slope variation; it represents the rate of slope change where the original signal curvature has been measured. Hence, this spectra pre-treatment will correct both baseline shift and drifts.

In this research study, the NIR spectra were smoothed first, followed by the second derivative with the third-order polynomial order using the Savitzky-Golay algorithm to correct baseline shift, drifts and random noise (Savitzky & Golay, 1964). There is no specific protocol for selecting how many smoothing points, either first or second derivative, and the polynomial order to be applied. Hence, it is entirely dependent on the chemometrician to evaluate further and decide which selection is the most appropriate to be applied accordingly. For this research study, both smoothing and derivative have been

used for FT-NIR spectra pre-treatment with SG second derivative, 17 smoothing points and polynomial order of three.

2.3.2 Multiplicative Scatter Correction (MSC)

MSC is another FT-NIR spectra pre-treatment technique which is commonly used. Initially, the MSC technique was developed for data normalisation for FT-NIR spectra amplification and baseline offset removal. Both amplification and baseline offset correction can be done via regression of the FT-NIR measured spectrum versus the reference spectrum. Subsequently corrected the FT-NIR measured spectrum using the slope of this fit (Martin et al., 1983; Geladi et al., 1985).

The algorithm equation of the MSC model for each spectrum is represented below in equation 2.1;

$$X_i k = a_i + b_i X_i + e_i k \; (i = 1 \ldots . N; \; k = 1 \ldots . k) \tag{2.1}$$

where *i* is the sample number and *k* is the wavelength or wavenumber. The constant a_i is the 'common shift' related to the proportional additive effect, while b_i represents the 'common amplification' and is related to the multiplicative effect for sample *i*.

The selection of FT-NIR spectra pre-treatment technique depends on the individual chemometrician to select the suitable technique to eliminate the spectra noise and drift prior to qualitative and quantitative modelling to be performed accordingly. Table 2.1 shows the types of NIR spectra pre-treatment and their purposes.

Table 2.1: Types of NIR Spectra Pre-Treatment

No.	Types of NIR Spectra Pre-Treatment	Purpose	Remarks
1	Savitzky-Golay Smoothing	Smooths the digital data points, increasing the data's precision without distorting the signal.	This technique is also known as convolution. Using the linear least-squares method, the technique will fit successive subsets of close data points with a low polynomial order.
2	First and second derivative	A baseline correction method. The constant background signals will be removed. The first derivative removes the constant offset.The second derivative removes the offset plus a linear term. The particle size can change the offset and spectra slope, requiring a second derivative technique.	Calculating the differences between absorbance levels at successive wavelengths is the primary method for derivatisation. The first and second SG derivative algorithm includes smoothing to limit the spectra noise increment with the identified polynomial order.
3	Multiplicative Scatter Correction (MSC)	MSC is a technique to correct the spectral signal noise and background effects. The light scattering or the change of cell path length contributes to spectral baseline shifting and tilting.	MSC uses the linear least-squares method to fit a linear model between a reference spectrum and other spectra of the dataset.

2.4 Chemometrics multivariate methodology

Chemometrics is used to solve problems involving a large amount of data. The extensive data are obtained from process control and blending operation with the chemical and physical univariate laboratory reference method analysis and FT-NIR spectroscopy spectra measurement. The extensive data obtained must be analyzed and visualized to evaluate the insight of the data gathered adequately for qualitative and quantitative measurements.

Chemometrics is a multidiscipline combination of chemistry, mathematics, statistics and common sense. It uses designing and selecting the optimal experimental procedures and providing applicable chemical information by analyzing chemical data and obtaining knowledge about the chemical system (Vandeginste et al., 1998).

Exploring the data will provide insight for qualitative measurement where the classification or clustering model can be developed by using different techniques such as Principal Component Analysis (PCA), Discrimination Analysis (DA), Hierarchical Cluster Analysis (HCA) and others.

For quantitative calibration models development, there are a few regression techniques that can be used such as Principal Component Regression (PCR), Partial Least Square Regression (PLSR) and Multiple Linear Regression (MLR). PLSR, PCR and MLR are multivariate analyses that allow extracting the chemical information by analyzing the full FT-NIR spectrum (Martens & Stark, 1991; Vandeginste et al., 1998). Multivariate regression techniques involve more than one independent variable (X variables) and may correspond to multiple response variables (Y variables) for calibration models regression.

The application of multivariate analysis in various technical fields or industries extensively covers chemical, pharmaceutical, petrochemical, foods, forensic and other industries. Hence the multivariate calibration models play an essential role (Faber & Rajko, 2007). Multivariate techniques explore big data sets by decomposing the complex data into simpler structures, improving the interpretation and extraction of the available information. The selected technique can be used to construct the qualitative and quantitative analysis after the FT-NIR spectra pre-treatment.

2.4.1 Qualitative measurement – Pattern Recognition / Clustering

Pattern Recognition in chemometrics is one of the most common techniques used in the analysis of petroleum hydrocarbon to determine the patterns and clustering, i.e., qualitative measurement whereby characterization of different petroleum hydrocarbon products can be determined. There are many methods commonly used for chemical pattern recognition.

Pattern recognition methods assign some output value or label (eg. types of petroleum hydrocarbon) based on the input values (independent variables, i.e., FT-NIR spectra absorbance for each 1.0 cm^{-1} wavenumber interval) according to some specific algorithm. Supervised and unsupervised learning procedures are required to categorize the pattern recognition by generating the output value for this research work, the type of petroleum hydrocarbon.

The supervised learning procedure generates a model from the calibration or training set, in which the sample sets are appropriately labelled with the correct output, eg. types of petroleum hydrocarbons (gasoline, diesel and kerosene).

In contrast, the unsupervised learning procedure involves the calibration of training set in which the samples are not labelled accordingly. It will attempt to identify the

characteristic pattern in the data that can determine the correct output value, i.e., types of petroleum product cluster or group from the new data inputs, FT-NIR spectra independent variables.

There are various techniques for pattern recognition to determine classification or clustering. These techniques include Principal Component Analysis (PCA), discrimination analyses such as partial least square discrimination analysis (PLS-DA) and others. PCA is the most common chemometrics tool addressing discriminant classification and class modelling techniques in this literature review.

A fundamental chemometric algorithm extensively used is PCA. When the transformed new set of uncorrelated (orthogonal) components is used as the independent variables for a least-square method, the methodology is called the Principal Components Regression (PCR) (Næs et al., 2002). The new independent variable (the PC) is a linear combination of the original variables that lie along the direction of maximum variance in the data set. To obtain other PCs, the data projection is continued until all the significant structures of the data are described by composing additional (orthogonal) PCs.

A set of FT-NIR spectra (for this research it is 801 data sets of X variables) and 300 samples (for this research study) can be expressed as a data matrix X (n x p), which is 801 x 300. The data matrix comprises two low-dimensional matrices: the score matrix (T) and transposed loading matrix (P). The data matrix X contains n values of absorbance at each p wavelength or wavenumbers (André et al., 1997).

$$X = TP + E \tag{2.2}$$

From Equation 2.2, E is the residual matrix that contains the unsystematic variation or noises. The scores are the new values of the FT-NIR spectra in the coordinate system defined by PCs. The loadings or eigenvectors are the links between the wavelength or

wavenumber of the *X* matrix and the principal component. An important feature of PCA is the graphic interface, i.e., a plot of score and loadings. A two-dimensional or three-dimensional scatter plot of the scores depicts the covariance between samples, providing a data overview. Clusters or groups of the objects (e.g., types of petroleum hydrocarbon) and outliers and hence the pattern and cluster are easily identified in the score plots. The score from a two-dimensional or three-dimensional scatter plot illustrates the variance between samples, providing a data overview. Patterns and clusters make clusters or groups of objects (eg. types of petroleum hydrocarbon) and outliers can be easily identified. These will reveal the expected and unexpected trend in the FT-NIR spectra data and insight concerning the variation of the composition and process or batch petroleum product blending.

2.4.2 Quantitative measurement

Multivariate regression analysis technique selection is vital before the FT-NIR can be used to perform the multivariate measurement and quantitatively predict the desired qualities for each petroleum product hydrocarbon. This is to assure that the prediction measurement of each chemical and physical properties of petroleum hydrocarbon by FT-NIR is comparative with the laboratory reference methods with the acceptable confidence level.

Quantitative analysis refers to an analysis in which the concentration or amount of an analyte may be expressed and estimated as a numerical value in proper units (Currie, 1995). Quantitative analysis requires the identification or qualification of the analyte for which numerical estimates are given. The quantitative analysis develops regression models which attempt to predict a quantity based on measurement of responses independent *X* variables (NIR spectra) and corresponding quantities dependent *Y* variables (laboratory reference data) on known petroleum product hydrocarbon. For this

study, 801 independent *X* variables were used from the NIR wavenumber from 4800 to 4000 cm^{-1}.

The chemometrics multivariate regression analysis typically consists of three essential steps which are are:

2.4.2.1 Calibration models development

The development of quantitative relation of multivariate calibration required the digitized spectra as a matrix *X* and the laboratory reference values as a matrix *Y* (Martens & Næs, 1989).

Multivariate calibration models can be developed by different regression techniques available. Multiple Linear Regression (MLR), Partial Least Square Regression (PLSR) and Principal Component Regression (PCR) are the most used techniques for linear regression methods. Other regression techniques which can be applied for non-linear information extraction from FT-NIR spectroscopic data, are Support Vector Machine Regression (SVMR), Locally Weighted Regression (LWR) and Artificial Neural Networks (ANNs). The PCR and PLSR techniques were used for calibration model development for this research study.

When contiguous variables are highly correlated, a covariance problem must be solved. PCR provides a simple solution. A traditional least-squares method is used to create the model, which relies on a smaller number of significant main components calculated from the original variables as the predictors. (Jolliffe, 1982). The PCs are uncorrelated since they are orthogonal by definition. According to PCR, the lowest order PCs, or those with the biggest variance, are considered to be the most crucial in predicting a response variable. A methodical selection of the PC to be included in the model based on their modelling efficiency is an important step.

The most popular multivariate regression method is undoubtedly partial least squares (PLS), which offers a superior answer for situations involving variable number and inter-correlation (Wold et al., 2001). PLS can also refer to projections onto latent structures. The latent structures are directions in the space of the predictors, also known as latent variables (LVs) or PLS components. The direction with the highest correlation with the chosen response variable is particularly indicative of the first latent variable.

The first latent variable's information is then eliminated from the response as well as the initial predictors. The direction of highest covariance between the residuals of the predictors and the responses is the second latent variable, which is orthogonal to the first. Same strategy is applied for subsequent LV. Evaluation of the prediction error corresponding to models with increasing complexity using an acceptable validation approach yields the optimal complexity of the PLS model, or the most suitable number of latent variables.

In summary, the advantages and disadvantages types of quantitative multivariate regression techniques (PLSR and PCR) are as follows:

- Experience in NIR data has shown that PLSR can give good prediction results with fewer components than PCR.
- A consequence of this is that the number of components needed for interpretation of the information in X which is related to y is smaller for PLSR than PCR which leads to simpler interpretation.
- From a computational point of view, PLSR is usually faster than PCR. For large data sets this aspect may be of some value.
- From a theoretical point of view, PCR is better understood than PLSR.

Table 2.2 summarizes the types of quantitative multivariate regression techniques.

There is no specific or standard protocol for developing the multivariate calibration model and model performance validation. For ‘simple’ calibration, 35 to 40 samples are the optimum size for calibration data sets (Williams, 1987). However, several hundreds of samples are required for ‘complex’ calibration, such as in forge calibration sets.

There are various ways used by chemometricians in developing the calibration models, covering the following but not limited to:

- Identify and select the calibration sets and validation sets, respectively, such as minimum numbers for each set, by date or batch, and others.
- Identify the outliers of the data, either independent variables (FT-NIR Spectra) or dependent variables (reference laboratory test data).
- Identify the outliers, leverage and residual of the data, independent and dependent variables.
- Identify or selection of the optimum principal components or latent variables
- Evaluate regression coefficient and determine the optimum principal components or latent variables comparative with Root Mean Square of Calibration Error.
- Avoid under-fitting or overfitting in multivariate calibration model development.
- Evaluate the calibration model's performance comparable with reference laboratory test methods.

Table 2.2: Types of Quantitative Multivariate Regression Techniques

No	Types	Algorithm	Method	Remarks
1	Multiple Linear Regression (MLR)	Linear	A method for linking the variations in a response variable (Y-variable) to the variations of several predictors (*X*-variables)[a]	Collinearity can lead to misinterpretation when using the MLR method. It is often the case for the sets of variables employed. The assumption is that the *X*-variables are linearly independent, meaning no linear relationship exists between *X* variables. The number of samples shall be higher than the *X*-variables at all times.
2	Principal Component Regression (PCR)	Linear	A two-stage operation in which the *X*-variables are first subjected to a standard principal components' decomposition exactly. Then the Y– variables regressed onto this decomposed *X* – matrix.[b]	PCR eliminates the collinearity and significant *X*- variables error, i.e., spectral. Not the best fit if modelling more than one Y-variable simultaneously caused PCR well decomposed for *X* – matrix but not optimal for Y- variables prediction.

No	Types	Algorithm	Method	Remarks
3	Partial Least Square Regression (PLSR)	Linear	Designed to cope fully multivariate regression case for both *X*- and Y- spaces for multivariate.[c] PLS can handle one or more of several co-varying Y-variables equally well. They were used as a supervised calibration tool that simultaneously classifies the new *X*-vectors submitted for prediction.	Handle the weaknesses of MLR and is an improvement over PCR in terms of prediction ability. Most generalised multivariate regression techniques, with complete control over both collinearity and *X*-errors.
4	Artificial Neural Network (ANN)	Non-Linear	An artificial neural network (ANN) is a computing system designed to simulate how the human brain	Artificial Intelligence (AI) is based on the ability to solve problems that would be impossible or difficult to solve using human or statistical methods. As more data becomes

39

No	Types	Algorithm	Method	Remarks
			analyses and processes information.[d] ANNs are made up of processing units, which have inputs and outputs. In order to produce the required output, the ANN learns from the inputs. The set of learning principles that artificial neural networks follow is called backpropagation.	available, ANNs has self-learning capabilities to deliver better outcomes.

[a] (Dongare et al., 2012; Pirhadi et al., 2015; Rosenfeld et al., 1998; Wold et al., 2001) [b] (Rosenfeld et al., 1998) [c] (Wold et al., 2001) [d] (Agatonovic-Kustrin & Beresford, 2000)

2.4.2.2 Calibration model validation and prediction.

The validation set is employed if the data collected or gathered is sufficiently large to be divided into calibration sets and validation sets, respectively. The validation set used for prediction ability, i.e., to estimate the prediction error between FT-NIR calibration models prediction, is comparable with laboratory reference test data or results.

Depending on the number of samples used for the validation, prediction ability values should be reported together with their respective confidence intervals. In every modeling procedure, estimating the predictive power on fresh samples that were not utilized to create the models is a crucial stage. For this, a number of procedures have been used.

A single validation set is the most uncomplicated and most rapid validation scheme. It has been divided typically approximately 1/3 of the total number from the calibration set. The subdivision could be random, arbitrary, or carried out using a uniform design, such as the Kennard and Stone and the duplex algorithm which enables the acquisition of two uniformly distributed subsets that are representative of the overall sample variability.

The most common validation protocol is the cross-validation procedure. The *N* available samples are distributed into *G* cancellation groups following a pre-determined scheme. The model is computed *G* times where each time, one of the cancellation groups is used as the validation set. At the end of the procedure, each sample has been used $G - 1$ time for building a model and once for an evaluation.

The leave-one-out procedure is a form of cross-validation with N cancellation groups. It is unique for a given data set and is generally known for its advantage, whereas different subdivision schemes and orders of the samples generally yield different outcomes when $G < N$.

2.4.2.3 Calibration model and validation performance evaluation

It is essential to evaluate the calibration model and validation performance prior to the multivariate measurement prediction by FT-NIR being technically valid, i.e., accurate and precise for real applications.

Several measures of determining the calibration model and validation performance are good. How successfully the A-dimensional model has been fitted to the calibration data set can be determined by either the model's fit or lack thereof. The prediction error expresses the error anticipated when using a calibration model to future predictions. The correlation between predicted and measured (reference laboratory method data) values is another way to evaluate the calibration model performance. The residuals reveal how well each object is modelled and predicted.

The quality of calibration and prediction equations is described using a variety of statistics. These are listed, with calculation methods, by (William, 1987) and summarized in Table 2.3 below.

Table 2.3: Statistic to describe the FT-NIR spectroscopy calibration and prediction equation quality.

Statistic	Definition
Standard Error of Calibration (SEC)	Derived from the equation developed from the calibration data set resulting in the difference values between the predicted value (FT-NIR) and the laboratory reference method.
Standard Error of Prediction (SEP)[a]	Derived from the equation applied to the validation data set, resulting in the difference between predicted values.

Table 2.3, continued.

Statistic	Definition
Standard Error of Cross-Validation (SECV)	Derived from the equation applied from a subset of the calibration data set resulting in the difference values between the predicted value (FT-NIR) and the laboratory reference method.
Coefficient of determination (R^2)	A statistical measure in a regression model determines the proportion of variance in the dependent variable that the independent variables can explain.
Correlation coefficient (r)	A statistical measure of the degree to which changes to one variable's value predict change to another's value.
Bias (D)	The average values of the difference between the predicted (FT-NIR) and laboratory reference test method.

[a] SEP is also can be defined as "standard deviation of performance", "standard error of estimate", standard error of analysis and standard error selection (Brown, 1990; Smith et al., 1991; Williams, 1987)

The derivation of the SEC by (Smith et al., 1991) is shown in equation 2.3;

$$SEC = \Sigma(X_i - Y_i)^2/(N - p - 1)^{0.5} \quad (2.3)$$

where X_i represents the predicted value of the i^{th} item in the validation set, Y_i is the reference value of i^{th} item in the validation set, N is the number of items in the validation set and p is the number of independent variables in the prediction equation. Referring to Adesogan et al.'s research, the equation that should be selected needs to involve the largest R^2, smaller SEC and lowest number of spectral terms to avoid overfitting (Adesogan et al., 1998).

The validation of the calibration model is tested or validated against another data set which has not been used in the calibration data set with reference values (laboratory reference data). The predicted values will typically differ from the reference values. This is because of random error. However, there are two types of systematic errors.

If the regression coefficient is different from 1.0, (Williams, 1987) noted that this would introduce a systematic bias at either end of the range of predicted values. It also may be contributed from the reference values. Hence, the standard error of prediction corrected for bias (SEP(C)) can be calculated as follows (Smith et al., 1991):

$$SEP(C)\Sigma(X_i - Y_i)^2 - N(bias)^2/(N-1)^{0.5} \quad (2.4)$$

From equation 2.4, X_i is the predicted value of the i^{th} item in the validation set, Y_i refers to the reference value of i^{th} item in the validation set, Bias is the difference between overall means and N is the number of items in the validation set. Williams, (1987) has provided rules for interpreting values from bias, SEP and correlation between predicted and reference values. He recommended that the SEP should not be more than 3% of the mean reference value for that analyte.

Shenk et al. (1985) recommended that the SEP should not be greater than twice the SEL (Standard Error of Laboratory) (Shenk et al., 1985). The size of the bias2: SEP2 ratio in proportion to the mean of the reference values can be used to measure bias (Park et al., 1998). This ratio should be small. By changing the regression intercept, a uniform bias can be eliminated. Predicted values can be replaced at either end of the reference range as described by Williams, (1987).

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Petroleum Products Sampling

Table 3.1 shows the number of samples that were collected for qualitative and quantitative measurements.

Table 3.1: Number of qualitative and quantitative sample

Type of measurement	Type of samples	Number of samples	Parameters	Test Method
Qualitative measurements	Gasoline	100	N/A	N/A
	Kerosene	100		
	Diesel	100		
	Diesel without PME blend	40		
	Diesel with PME blend	36		
	Gasoline without additive	40		
	Gasoline with additive	44		
Quantitative measurements	Diesel	266	Boiling point at T95 recovery	ASTM D86-20 (2020)
			Flash point	ASTM DD93-20 (2020)
			Could point	ASTM D2500-17 (2017)
			Calculated cetane index	ASTM D976-06 (2016)

3.1.1 Qualitative measurement samples

One hundred (100) of each gasoline, kerosene and diesel samples, including directly from process streams, during blending and final product tanks were collected from different processing units and blending systems (different times, dates and batches) between December 2019 and August 2020. Collected samples were from oil and gas process plants locally which was in Malaysia. During sample collection, the product must be kept idle for a minimum of 2 hours to ensure the separation of oil and water in the tank. Water which has a higher density will settle down and drain from the tank leaving pure oil content. The sample were then stirred homogenously for 1 hour before sample was collected.

Forty (40) of diesel without PME blend and thirty-six (36) diesel with PME blend samples were collected from refinery and petrol station, respectively. The same location of sampling station was used to ensure consistency of results and minimize error. These samples were pumped and placed in a high density poly-eutherene (HDPE) container for safety purposes.

Forty (40) of gasoline without additive and forty-four (44) gasoline with additive samples were collected from the refinery and petrol station. The same method of collection was applied to all samples. Safety gloves were also used in the process to minimize contamination of samples.

All the samples were stored in an amber glass bottles pending FT-NIR spectra measurement. The samples were stored in a chiller for gasoline and kerosene to avoid vaporising lighter components, leading to sample integrity. For diesel, samples were stored and kept in the darkroom at room temperature to avoid any chemical reaction when exposed to light.

3.1.2 Quantitative measurement sample

Two hundred sixty-six (266) diesel samples (including single and blended components) were collected from three different processing units (different times, dates and batches) in Malaysia between December 2019 and August 2020. The samples were stored in amber glass bottles and kept in the darkroom while pending laboratory analyses.

3.2 Determination of Boiling Point at 95% Recovery

Based on samples composition, vapour pressure, and the expected initial boiling point (IBP) and final boiling point (FBP) or expected endpoint (EP), they belong to Group 4 diesel fuel as stipulated in the test method ASTM D86-20 (2020). The apparatus arrangement, condenser temperature, and other operational variables are configured in the equipment operating system as described in Table 3.2.

Table 3.2: Operating settings for the measurement of boiling point

Parameters	Settings
Flask, mL	125
ASTM distillation thermometer	8ºC (8F)
IP distillation thermometer range	High
Flask support board	Type C
Diameter of the hole, mm	50
Temperature flask at start of the test ºC	Ambient
Receiving sample and sample, ºC	Ambient

A 100 mL diesel sample was distilled under prescribed conditions for group 4. The distillation was performed in a laboratory automated distillation unit at ambient pressure. The design of the distillation unit is approximately one theoretical plate fractionation (Figure 3.1). Systematic observations of temperature readings and condensate volumes were made until it

reached the final boiling point volume % recovery. At the end of the analysis, the volume of the residue and the losses were recorded.

At the end of the distillation analysis, the observed vapour temperatures at ambient pressure were corrected to atmospheric pressure by the built-in barometric pressure and software conversion. The final test results were expressed as per cent recovered versus the corresponding temperature, i.e., boiling point temperature at 95% recovery (T95).

Figure 3.1: Automated distillation unit

3.3 Determination of Flash Point

Test method ASTM DD93-20 (2020) by automated Pensky-Martens Closed Cup (PMCC) was used to determine the flash point of diesel samples (typical diesel flash points is in the range 50 ^{o}C to 80 ^{o}C). Procedure A was used for diesel fuel samples. The PMCC tester is illustrated in Figure 3.2.

The automated apparatus was designed and capable of performing the procedure described in the test method, with the proper heating rate control, the diesel samples' stirring, applying the ignition source, detecting the flash point, and recording the final flash point result. At least 75 mL of diesel sample was filled into the brass test cup and tested according to the settings shown in Table 3.3.

Table 3.3: Operating condition of automated Pensky-Martens closed cup flash point tester

Parameters	Requirements
Temperature of test cup and diesel sample, ºC	At least 18 ºC below the expected flash point.
Flame diameter, mm	2.2 to 4.8
Heating rate, ºC/min	5 to 6
Stirring rate, rpm	90 to 120
Ignition source application	23 ºC (± 5 ºC) below the expected flash point and each time after that at a temperature reading that is a multiple of 1 ºC.

The ignition source was directed into the test cup at regular intervals with simultaneous interruption of the stirring until a flash was detected. The observed flash point was reported to the integer number and corrected to atmospheric pressure.

Figure 3.2: Automated PMCC flash point tester

3.4 Determination of Cloud Point

The cloud points of diesel samples were measured according to the procedures described in ASTM D2500-17 (2017). This test method was designed to determine diesel fuels cloud points below 49 ºC by using a manual cooling bath, as illustrated in Figure 3.3.

The diesel samples were cooled for at least 14ºC above the expected cloud point. Diesel samples were then filtered to remove any moisture present using dry lint less filter paper until oil appears on the filter paper. The diesel samples were poured into the test jar to the level mark and using a high cloud point ASTM thermometer to monitor the temperature. The test jar was closed tightly by the cork inserted together with the test thermometer. The thermometer bulb

or probe was rested on the bottom of the jar. The cork and the thermometer were adjusted to ensure the thermometric device and the jar were coaxial.

Figure 3.3: Manual cloud point apparatus and cooling bath

The disk was then cleaned and dried. It was placed on the bottom of the jacket in the cooling medium for at least 10 minutes prior to inserting it into the test jar. The gasket was then placed around the test jar about 25 mm from the bottom, and the test jar was inserted into the jacket.

Inspection of diesel samples cloud point was conducted at a multiple of 1 °C thermometer readings by removing the test jar from the jacket quickly but without disturbing the diesel samples and replacing it in the jacket, not more than 3 seconds. The cooling bath temperature was maintained at 0 °C ± 1.5 °C. If the diesel samples do not show a cloud when it has been cooled to 9 °C, transfer the test jar to a jacket in a second bath maintained at a temperature of −18°C ± 1.5°C (see Table 3.4).

If the diesel sample does not show a cloud when it has been cooled to −6 °C, transfer the test jar to a jacket in a third bath maintained at a temperature of −33°C ± 1.5 °C as described in Table 3.4 until the diesel sample exhibited a cloud point.

Table 3.4: Bath and sample temperature ranges

Bath	Bath temperature setting, °C	Sample temperature range, °C
1	0 ±1.5	Start to 9
2	-18 ±1.5	9 to -6
3	-33 ±1.5	-6 to -24
4	-51 ±1.5	-24 to -42
5	-69 ±1.5	-42 to -60

The cloud observed at the bottom of the test jar was reported to the nearest 1 °C, as the cloud point of the diesel samples tested.

3.5 Calculated Cetane Index

The cetane index of each diesel sample was calculated according to ASTM D976-06 (2016). This method covers the Calculated Cetane Index (CCI) formula, representing a means for directly estimating the ASTM cetane number of diesel fuels. The CCI is derived from American Petroleum Institute (API) gravity or density at 15 °C and the boiling point temperature at 50 % recovery. The index value is termed the CCI and computed from the formula given in Equation 3.1.

$$CI = 454.74 - 1641.461D + 774.74D^2 - 0.554B + 97.803(\log B)^2 \quad (3.1)$$

where D is the fuel density at 15 °C (ASTM D1298-12b, 2017) and B is the boiling temperature at 50 % recovery derived from the distillation curve (ASTM D86-20, 2020). Figure 3.4 displays the analysers used for physical distillation and densitometer.

Figure 3.4: a) Physical distillation and b) densitometer analysers.

3.6 Fourier Transform-Near Infrared Spectrometry

FT-NIR spectra were obtained under the condition specified in Table 3.5 using an ABB MB 3600 Series Laboratory FT-NIR spectrometer and Horizon MB software. Horizon MB is for FT-NIR spectra measurement whereas Horizon QA is used for sample measurement or analysis. Immersion probe with an optical path of 0.5 mm and a transmittance sample cell of CaF_2 beam splitter were used in combination with Indium Arsenide (InAs) detector as illustrated in Figure 3.5. Before each measurement, the immersion probe was cleaned with spectroscopic grade n-pentane and flushed with the samples to be measured. The volume of each sample was approximately 10 mL.

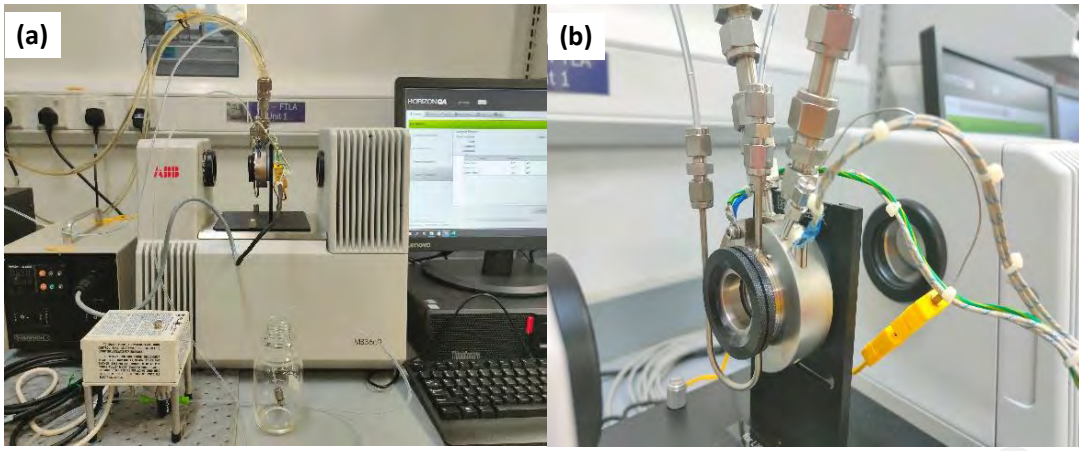


Figure 3.5: (a) ABB MB 3600 Series Laboratory FT-NIR spectrometer (b) Harrick cell CaF_2 complete with optical and thermostat probes.

Table 3.5: ABB MB 3600 FT-NIR operating condition

Parameters	Settings
Cell pathlength, mm	0.5
Cell measurement temperature, ^oC	25.0 ±1.0
Wavenumber, cm^{-1}	3,700-14,700
Number of scans	32
Resolution interval, cm^{-1}	1.0
Detector	InAs

3.7 Multivariate Data Analysis

Two independent datasets were organised and sorted into spreadsheet for qualitative and quantitative analyses.

The experimental data for qualitative evaluation was organized into a 300×801 matrix (100 gasoline, 100 kerosene and 100 diesel samples, with 801 spectral variables). The data were randomly divided into calibration (80 samples for each gasoline, kerosene and diesel) for PCA

model development and validation sets (20 samples for each gasoline, kerosene and diesel) for model validation purposes, respectively.

For eighty-four (84) gasoline with and without additive and seventy-six (76) diesel with and without PME blend, the data were organized into an 84×801 and 76×801 matrix respectively.

For quantitative study, the data was organized into a 266×806 matrix (266 diesel samples, five physio-chemical variables, 801 spectral variables). The dataset was randomly sectioned into a calibration set (177×806) and a validation set (89×806) for model calibration and validation purposes.

Subsequent spectral pre-processing and data analyses were performed with the Unscrambler® 9.8 by Camo Software.

3.7.1 Multiplicative scattering correction

MSC is a transformation method that can be used to compensate for both additive and multiplicative effects. In MSC, the corrected response of FT-NIR is given by the expression shown in equation 3.2.

$$x_{ij}^{*} = \frac{x_{ij} - b_1}{b_0} \tag{3.2}$$

Where x_{ij} is the original response of i-th diesel sample at j-th wavenumber; b_1 and b_0 are the intercepts and slope coefficients estimated by regression of the sample spectrum against the mean spectrum derived from the training set.

3.7.2 Savitzky-Golay derivatisation

SG-SD spectrum was obtained from a sequential local fitting and numerical derivatization (Savitzky & Golay, 1964). In this study, the FT-NIR response at a window size of eleven points was fitted onto a third-order polynomial, and the second derivative of the centre point was extracted. This process was propagated to subsequent windows to cover the remaining range of wavenumber.

3.7.3 Principal component analysis

PCA models were constructed by transforming the original data of correlated variables (independent variables *X* and samples *N*) to a new reduced set of orthogonal variables, the principal components (PCs), each containing unique information (Jolliffe & Cadima, 2016). The true distances are only exact in the plot if the PCs shown explain 100% of *X* variance.

3.7.4 Soft independent modelling of class analogy

Soft Independent Modelling of Class Analogy (SIMCA) was adopted to assign future samples based their residual distances from each class modelled with PCA using independent training set (Maesschalck et al., 1999).

3.7.5 Principal component regression

The PCR models were constructed by stepwise regression of each physio-chemical variable on a set of uncorrelated PCs (Frank & Friedman, 1993). These components were extracted using the Non-Linear Iterative Partial Least Squares (NIPALS) algorithm such that they explained maximal variance observed in the spectral variables (Jolliffe, 1982).

3.7.6 Partial least squares regression

Similarly, the PLSR models were also computed using the NIPALS algorithm. In this case, the resulting components described maximal covariance between the physio-chemical variables and the spectral variables (Frank & Friedman, 1993; Wold et al., 2001).

3.7.7 Model optimisation and validation

The optimum number of latent variables (LVs) in each model was estimated by considering together the root mean square error from leave-one-out cross-validation (RMSECV) with mean center and also the noise modelled by respective regression coefficient (ASTM E1655-00, 2000) via the Unscrambler 9.8 ® by Camo Software. The performance of the regression models was evaluated with the coefficient of determination (R^2), Residual Predictive Deviation (RPD), and root means square error of prediction (RMSEP) derived from the validation set and compared against the ASTM specification using the following equations:

$$R^2 = 1 - \frac{\Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2} \tag{3.3}$$

Where Y_i is the actual Y value, $\hat{Y}$ is the predicted Y value and $\bar{Y} = \frac{\Sigma Y}{N}$.

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}} \tag{3.4}$$

$$SEP(C) = \Sigma(X_i - Y_i)^2 - \frac{N(bias)^2}{(N-1)^{0.5}} \tag{3.5}$$

$$RPD = \frac{\sqrt{\frac{\Sigma(X_i - \bar{X})^2}{N-1}}}{RMSEP} \tag{3.6}$$

3.8 The conceptual framework

Figure 3.6 and 3.7 summarizes the overall flow used for qualitative and quantitative analysis respectively.

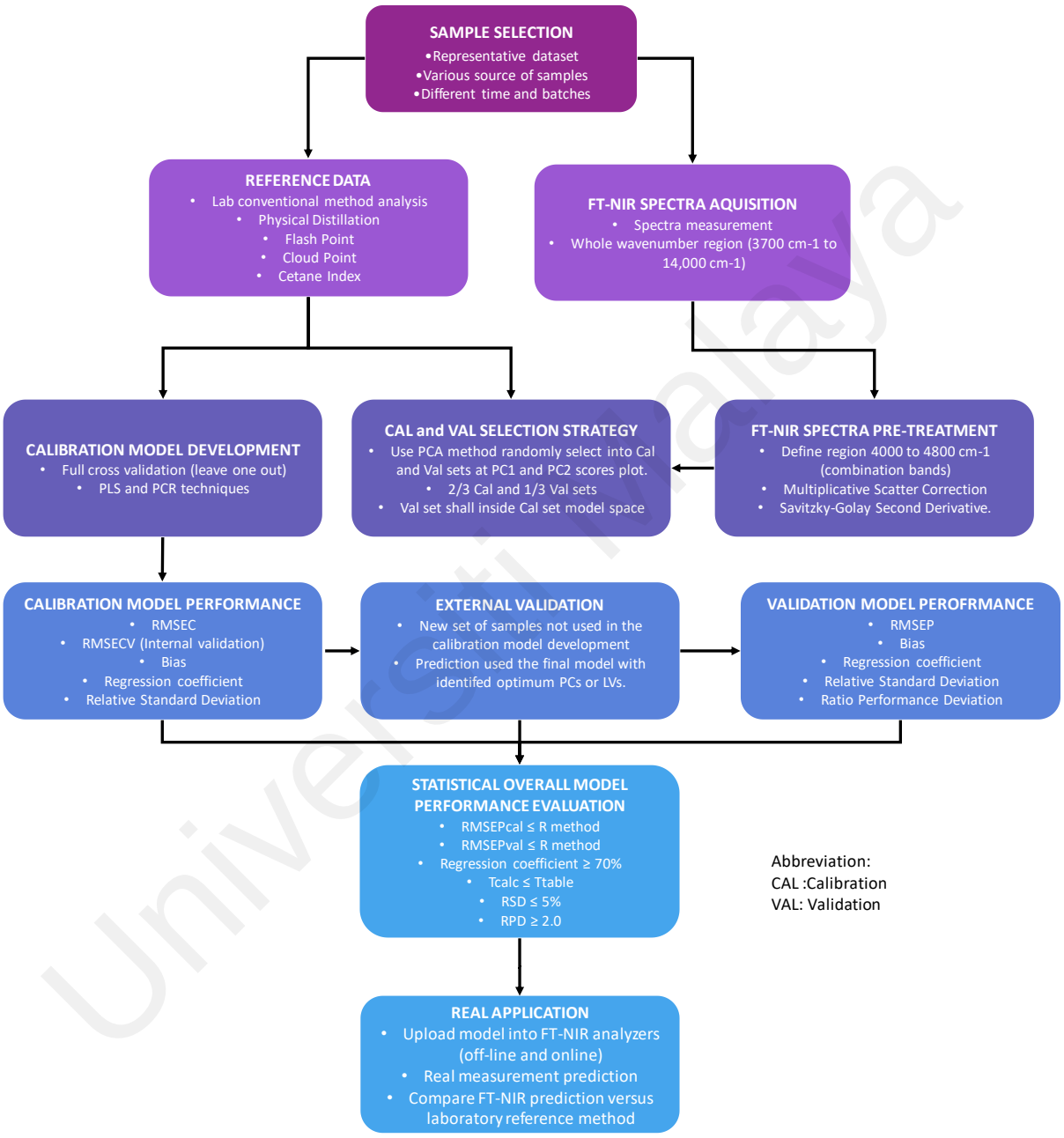


Figure 3.6: Conceptual framework for quantitative measurement.

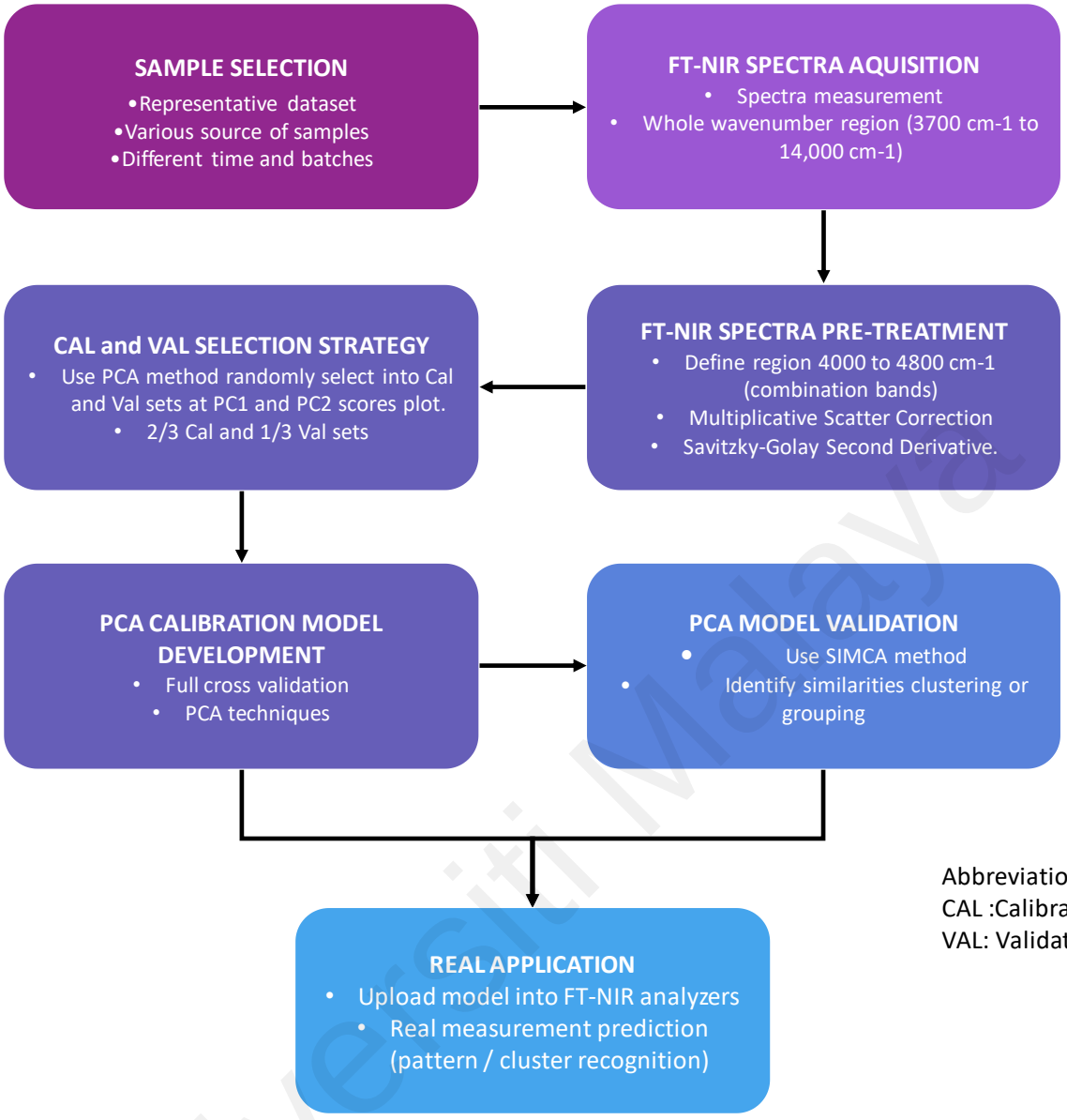


Figure 3.7: Conceptual framework for qualitative measurement

CHAPTER 4: RESULTS AND DISCUSSION

The fitness of a NIR multivariate model to predict physico-chemical properties of diesel samples relies on several factors, for instance, spectral region selection, signal pre-processing algorithm, calibration/validation set partition, and multivariate calibration strategy. More importantly, the predictive model must account for the batch-to-batch variations associated with diverse sources to serve the routine quality control/monitoring purposes. Table 4.1 summarizes the diesel samples variability observed in T95, FP, CP, and CI.

Table 4.1: Statistics for the measured properties of diesel samples

Properties	Mean	Sd Dev	Min	Max	Method	Reproducibility, %
T95 /°C	344.19	22.66	305.10	376.33	ASTM D86-20 (2020)	7.5
FP /°C	76.65	6.38	59.00	94.99	ASTM D93-20 (2020)	5.6
CP /°C	-6.31	7.06	-21.30	13.90	ASTM D2500-17 (2017)	4
CI	55.38	3.64	44.90	62.78	ASTM D976-06 (2006)	2

**As stated in the reference method*

4.1 FT-NIR Pre-Processing

To a great extent, the physico-chemical characteristics of diesel reflect the degree of refinement, degradation, oxidation etc., where the refined products vary from a clear thin liquid to thick opaque attributed to their complex hydrocarbon composition (Santana et al., 2007). Hence, selecting descriptive spectral variables that project the intrinsic configuration of diesel blends would provide vital

information enabling simultaneous determination of fuel quality parameters via the means of chemometrics. Considering that diesel composes mostly of paraffins, cycloalkanes, and aromatic compounds, the NIR responses at combination overtone region around 4800-4000 cm^{-1} that originated from harmonic vibrations such as methylene C-H, methyl C-H, O-H stretching and aromatic C-H vibrations etc. not only informative for Distillation, Flash Point, Benzene, Aromatics, Cetane Index, Freezing Point and other properties for calibration purposes, but also have avoided the unwanted spectral variations due to water content, fluorescence and scattering effects associated with the diesel samples (Alves et al., 2012).

Figure 4.1 displays the trimmed NIR spectra of the petroleum product samples before and after the signal pre-processing. It is expected that such pre-processing will improve the predictive performance of the calibration models by eliminating the artefacts introduced during the NIR measurement process (Mishra et al., 2020). Apparently, the multiplicative and scattering effects that dominated the variations observed in the raw spectra were managed by MSC, while SG-SD enhanced the spectral differences by addressing the random noise, baseline shift and drifts, which are either associated with the FT-NIR measurement method or the nature of diesel sample.

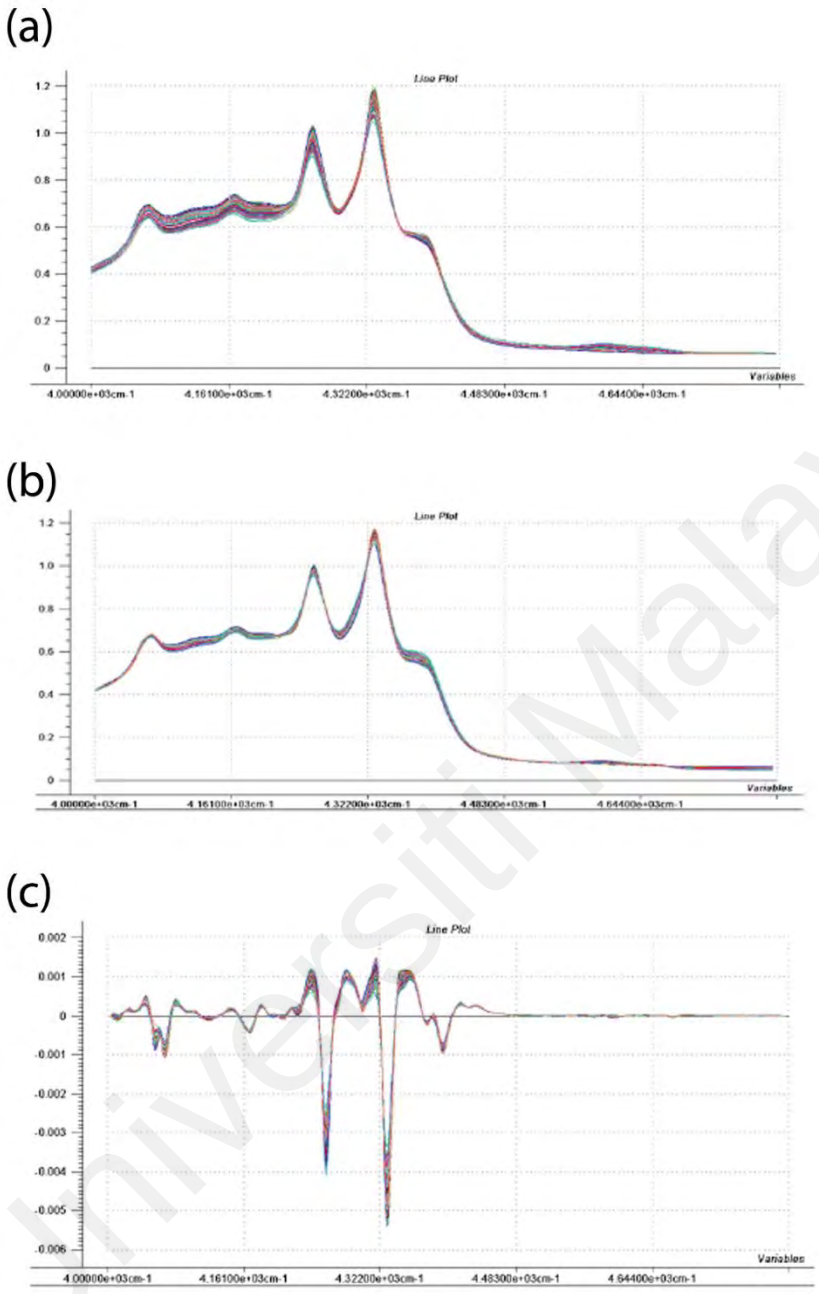


Figure 4.1: Near infrared spectra of the diesel samples: (a) raw spectra, (b) multiplicative-scattering corrected spectra, and (c) Savitzky-Golay derivatised spectra.

4.2 Multivariate Calibration

For robust modelling, the diesel samples were subdivided into calibration and validation sets based on a random sampling strategy to account for the batch-to-batch variances observed in routine samples. In this context, Figure 4.2 depicts the respective sample (spectral and physico-chemical) variabilities in a reduced space defined by the first two principal components. Looking at the dispersion of scores, it appears that such unbiased selection has yielded a good set of calibration points that covers the entire range of variation that typically encountered in routine quality control assessments and a comparable validation set that spanned over the calibration range. These score plots addressed the concern of representativeness of calibration and validation sets which has been voiced in previous calibration works (Liu et al., 2022; Palou et al., 2017).

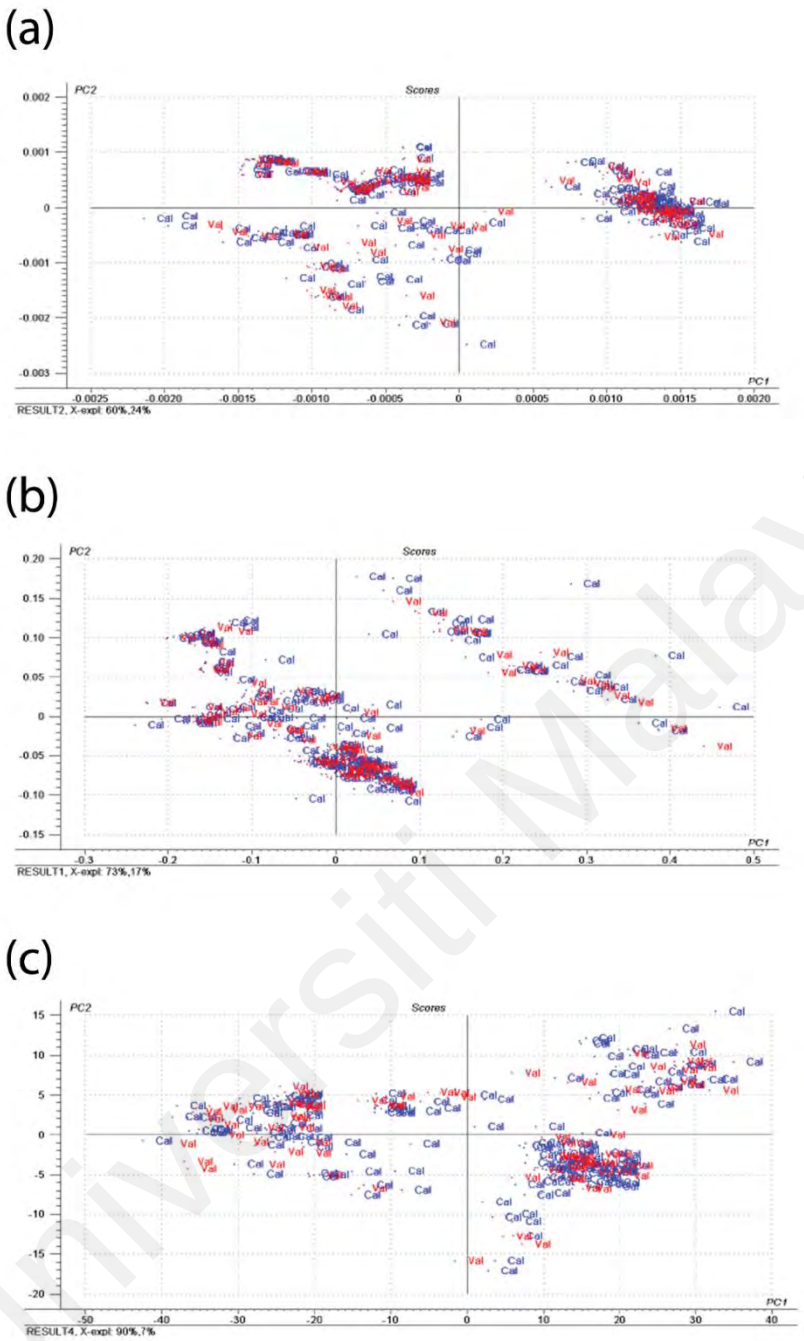


Figure 4.2: Principal component analysis scores plots of the diesel samples: (a) multiplicative-scattering corrected spectra, (b) Savitzky-Golay derivatised spectra, and (c) physico-chemical properties; where •Cal denotes the calibration samples and •Val denotes the validation samples

4.3 Qualitative Measurement – Grouping and Clustering Recognition

Principal Component Analysis (PCA) model was constructed from 80 samples from each petroleum products (Gasoline, Kerosene and Diesel) as a calibration set and 20 samples as a validation set, respectively. All the samples were measured by FT-NIR for spectrum data acquisition and pre-processed (MSC) prior to the model construction, as illustrated in Figure 4.3.

The hydrocarbon groups and compositions lead to differences in NIR spectra demonstrating the characteristic of each type of petroleum product. There are two main hydrocarbon groups namely Parffins (Aliphatic Hydrocarbons) and Aromatics will most differentiate the composition of gasoline, kerosene and diesel.

The two strongest bonds near 4330 cm^{-1} and 4259 cm^{-1} have been assigned to the symmetric and asymmetric modes of the combination of CH stretch and CH_2 bending motions respectively which represent aliphatic hydrocarbon groups. Gasoline has the highest concentration of aliphatic hydrocarbon groups compared to kerosene and diesel where diesel is the lowest.

At region 4166 – 4125 cm^{-1} the combination of CH stretching modes also clearly differentiated each of the petroleum products.

As compared with the benzene peak near 4058 - 4036 cm^{-1}, two bands are observed in the mono-, di- and trisubstituted methyl benzenes. At this NIR absorption bands differentiated between diesel with kerosene and gasoline, where diesel has two bands whereas kerosene and gasoline only have one absorption band. This is due to the diesel compositions having mono- di- and tri-aromatics hydrocarbon groups.

Kerosene and gasoline have only a single band that could be observed in higher substituted methyl benzenes at 4058 cm^{-1} corresponding to the CH_3 aromatic of symmetric bending vibrations. A linear relationship is found between the number of methyl groups substituted into the benzene ring and the intensity of characteristic absorption bands where diesel is the highest concentration and gasoline is the lowest.

The aromatic CH stretch produces several bands at a shorter wavelength than the aliphatic CH absorptions. Figure 4.3 for benzene the major NIR absorption bands are at 4668 – 4581 cm^{-1}. It is a very clear differentiation between the petroleum products which diesel has the highest aromatics absorption bands followed by kerosene and gasoline is the lowest.

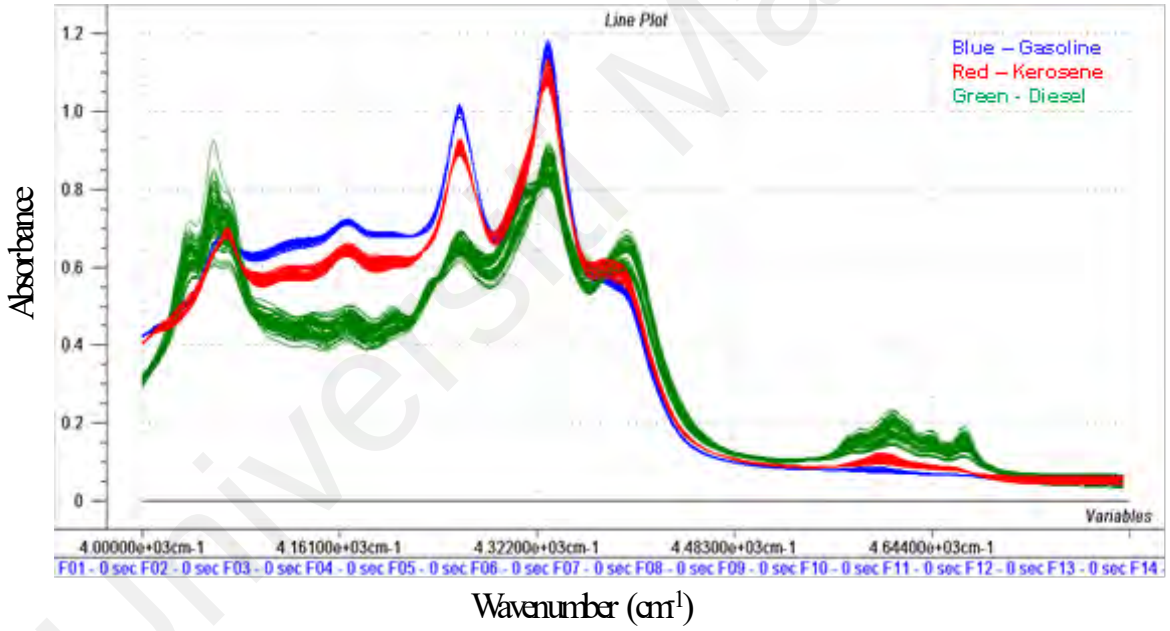


Figure 4.3: FT-NIR spectrum at combination band region 4800-4000 cm^{-1}

Figure 4.4 clearly shows that three petroleum product groups (Gasoline, Kerosene and Diesel) were recognized and well separated at PC1 and PC2, with the total *X* variance explained about 98%.

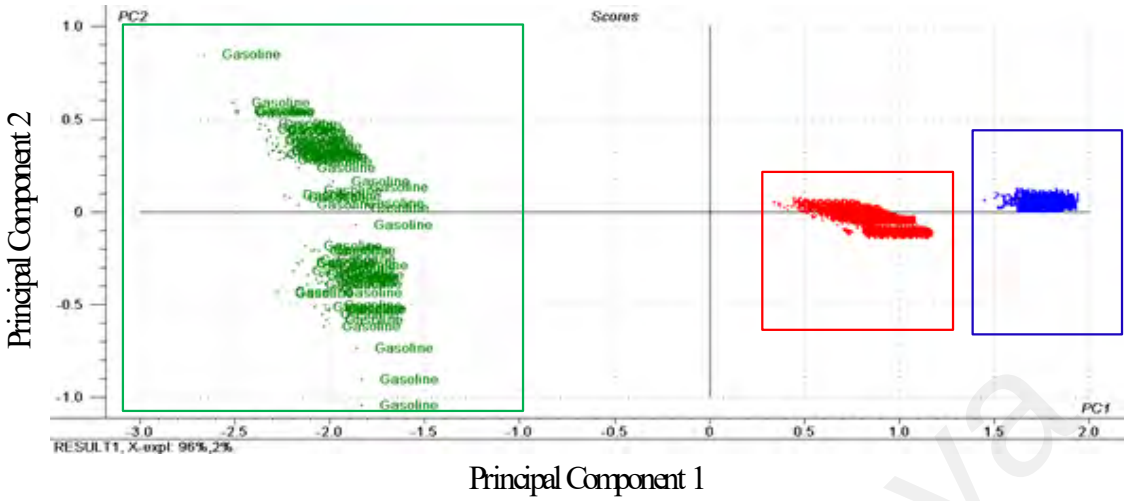


Figure 4.4: PCA model scores plot at PC1 and PC2 Legend: Green – Gasoline, Red – Kerosene, Blue – Diesel

Soft Independent Modeling of Class Analogy (SIMCA) classification was applied for qualitative validation steps. It focuses on modelling the similarities between members of the same class. A new sample will be recognised as a class member based on the similarities to the other class members; otherwise, it will be rejected. Twenty new samples for each petroleum product were validated based on the new FT-NIR spectra and predicted from the developed PCA model.

Chung et al., (1999) used FT-NIR spectroscopy to discriminate six types of petroleum products. The combination between PCA and Mahalanobis distance, products with similar properties and compositions such as light gasoil and diesel were efficiently identified with the accuracy of 99% (Chung et al., 1999; Li et al., 2018).

Figure 4.5 exhibits that all the new validation samples had been recognised and identified with the right group or cluster accordingly without the physical and chemical laboratory analysis, which is time-consuming and expensive. Note: where •C denotes the calibration samples, and •V denotes the new validation samples.

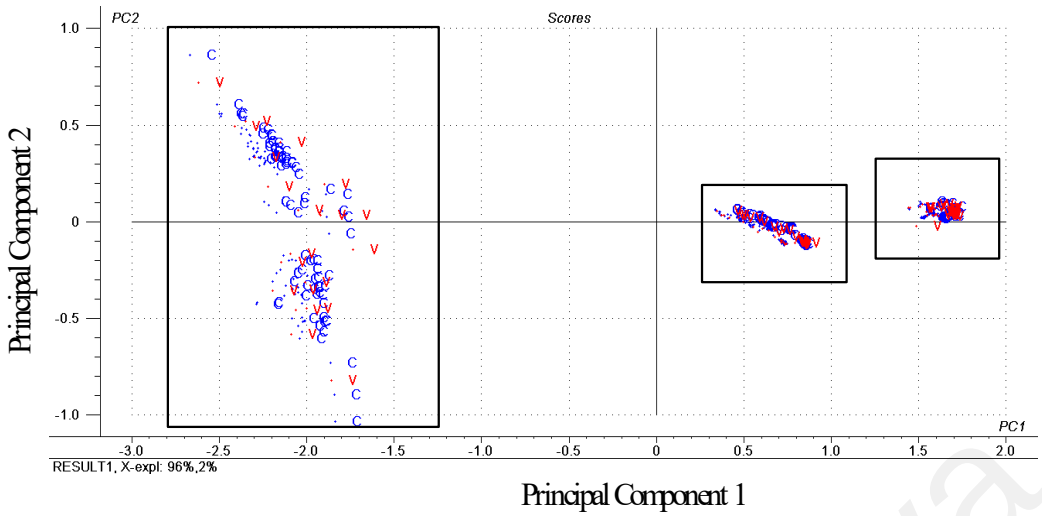


Figure 4.5: PCA qualitative model validation – SIMCA Clustering

Multivariate qualitative measurement is an alternative method that is rapid analysis, non-destructive, cheaper and able to reveal the insight information of the characteristic and fingerprinting types of the petroleum hydrocarbon products. It is beneficial for plant operators to pre-determine the cause of any contamination that occurred and product adulteration prior to delivery to the customers.

Skrobot et al., (2007) Chemometric data analysis was applied to investigate the gasoline adulteration at the gas station. The chromatographic data were used to identify the presence of solvents in gasoline. Using PCA, sample distribution patterns were investigated, and classification models were created with linear discriminant analysis (LDA). The results indicated the presence of solvent in gasoline effectively (Skrobot et al., 2007).

4.4 Qualitative Measurement for Gasoline with Additive and without Additive

Generally, the final blended gasoline product produced by refineries is without additives. It will be transported via tanker, vessel or pipeline into the terminal tank as the final destination, where additives will be added and mixed to improve the gasoline engine performance and emission. The typical concentration of additives added into gasoline in the range of 100 to 5000 ppm (wt) depends on the oil producers.

The main disadvantage of near-infrared is the low detection limit for the concentration of properties in ppm level. It is less accurate to determine these elements directly in gasoline using near-infrared spectroscopy. However, there is a study done to differentiate the gasoline with and without additives using near-infrared spectroscopy at the combination region, i.e. 4600 cm^{-1} to 4000 cm^{-1} (Silva et al., 2013).

From Figure 4.6, the spectra demonstrated slight differentiation within gasoline with and without additives in the region of 4100 cm^{-1} - 4000 cm^{-1} and 4700 cm^{-1} - 4500 cm^{-1}. However, the other regions, it had overlapped each other.

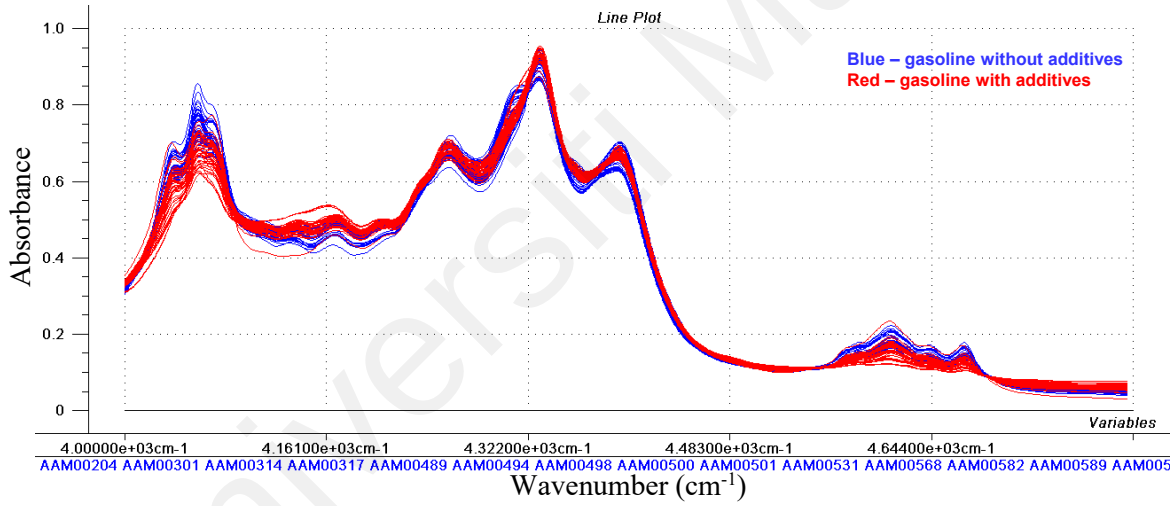


Figure 4.6: MSC treated spectra at 4800 cm^{-1} to 4000 cm^{-1} for gasoline with and without additives recorded by FT-NIR

For qualitative measurement, PCA was constructed as an exploratory to obtain more insight into the information recorded using FTNIR. Figure 4.9 demonstrated that total 90% of *X* variables had been explained at PC1 and PC2 where at PC1 and PC2 explained 72% and 18%, respectively.

Although at PC1 and PC2 score plots did not show clear separation or classification between gasoline with and without additives, it can observe some differentiation between them (see figure 4.7). This observation was expected since the additives added to the gasoline are in low concentration compared to the gasoline compositions, which are significant variations in concentration and compositions. There are two clusters or groups of gasoline without additives due to varying compositions from different petrol stations. A similar observation was observed at PC3 and PC4; with total *X* variables explained another 8%. Total *X* variables explained more than 95% from the sum of PC1, PC2 and PC3 as shown in Figure 4.8.

From the PCA above, the outcomes of the results show that the separation of the classes is not so clearly differentiated. However, at least it has shown a tendency of separation and provides some information about gasoline with and without additives.

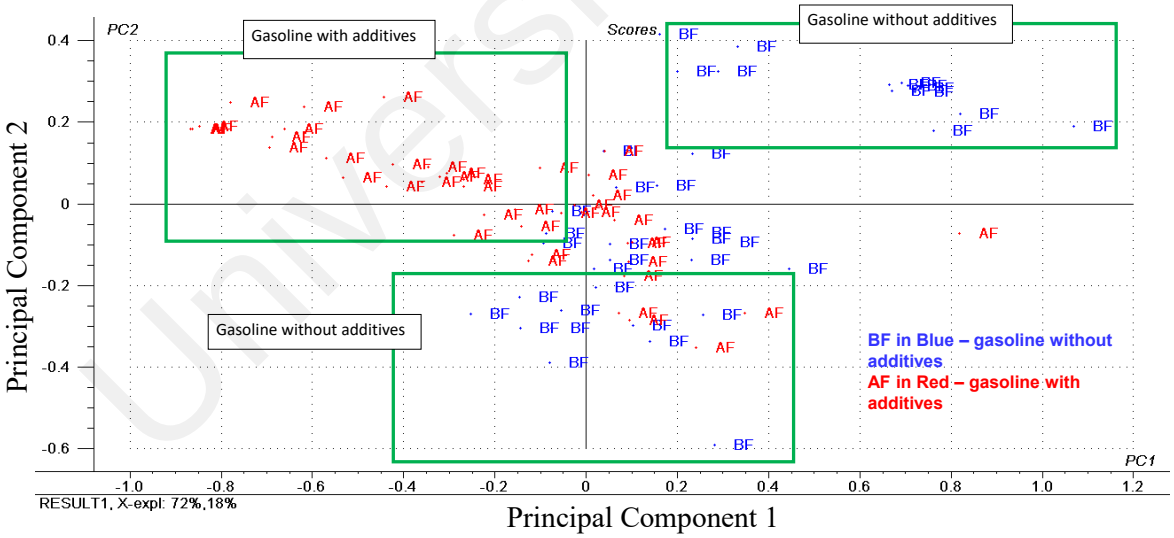


Figure 4.7: Scores plot at PC1 and PC2 derived from Principal Component Analysis

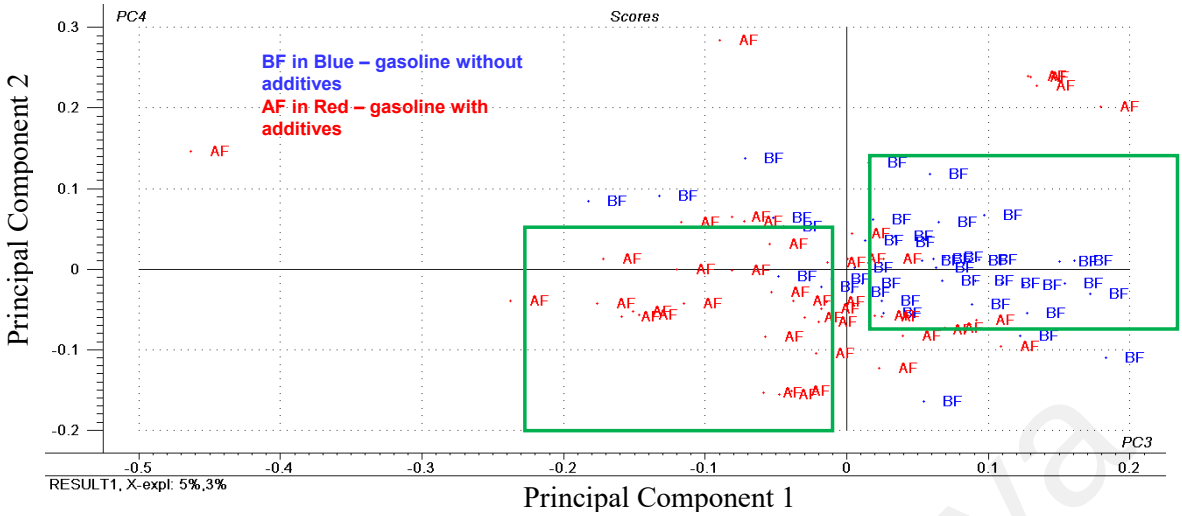


Figure 4.8: Scores plot PC3 and PC4 derived from Principal Component Analysis

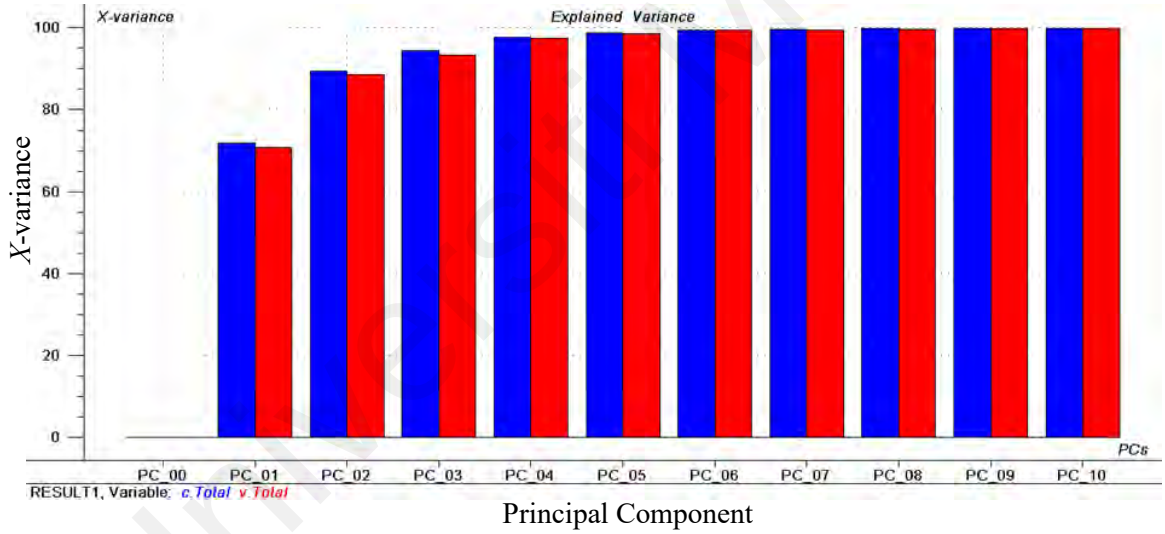


Figure 4.9: Explained variance for *X* variables plot

4.5 Qualitative Measurement for Diesel with and without blended with Palm Methyl Ester (PME)

Biodiesel is a renewable energy source that is less polluting than petroleum diesel and often blended with petroleum diesel (Lira et al., 2010). In Malaysia, the PMEs derived from palm oil

via the esterification process are a major component found in biodiesel. Malaysia's government has gazetted the biodiesel with the range of 7.1 to 20.0 volume/volume % concentration as specified in the Malaysia Standard High PME Diesel Fuel Specification – Euro 5 (MS 123-5, 2020). It is a national government agenda to support a zero-carbon emission declaration.

The main advantage of near infrared is a high detection limit for the concentration of PME in % volume level. It is accurate to determine and differentiate between diesel with and without blend with PME at the combination region, i.e. 4800 cm^{-1} to 4000 cm^{-1}.

FT-NIR spectroscopy combined with chemometrics multivariate methodology was used for this qualitative measurement study.

From Figure 4.10, the spectra demonstrated significant differentiation within diesel with and without blended with PME in 4081.1 cm^{-1} - 4049.0 cm^{-1} and 4452.0 cm^{-1} to 4424.0 cm^{-1}. As illustrated in Figures 4.11 and 4.12 (expansion from FT-NIR spectrum A and B), clear differentiation was observed. These combination regions the bands involving C=O and C–H. In diesel without PME, the C-H group concentration is higher than diesel with PME whereas the methyl ester group in diesel with PME is higher than in diesel without PME due to the diesel was blended with PME. However, the other regions, it had overlapped each other.

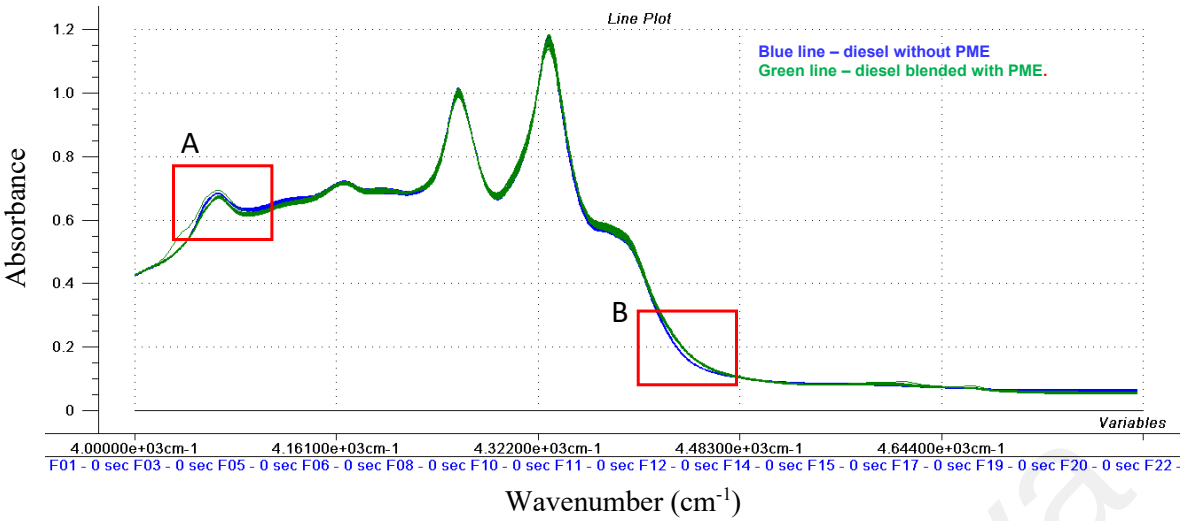


Figure 4.10: MSC treated spectra at 4800 cm^{-1} to 4000 cm^{-1} for diesel with and without blended with PME recorded by FT-NIR.

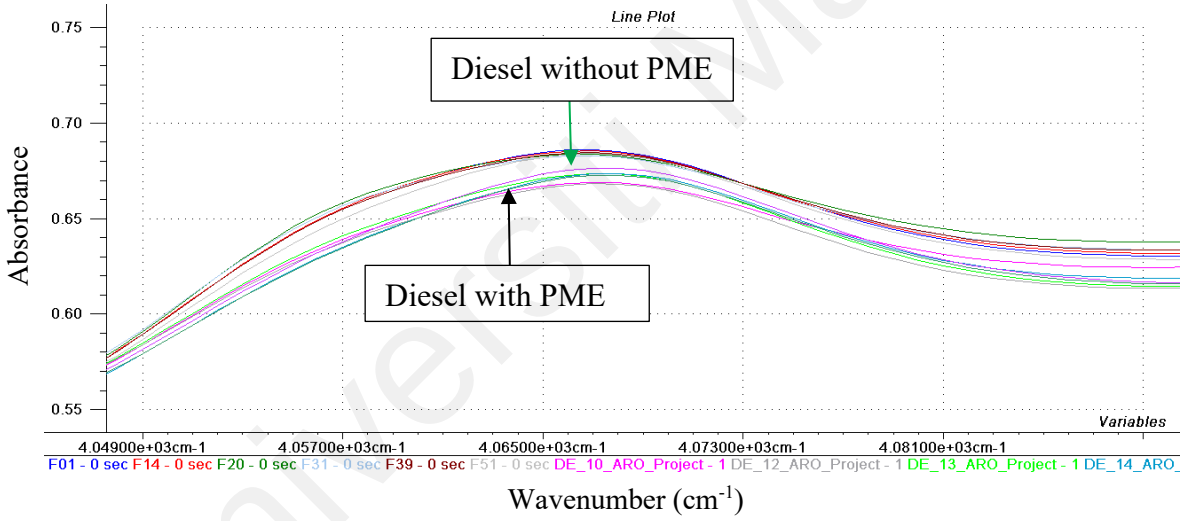


Figure 4.11: Spectra of diesel with and without PME blends between 4081.1 cm^{-1} to 4049.0 cm^{-1} (C-H group) combination overtone region

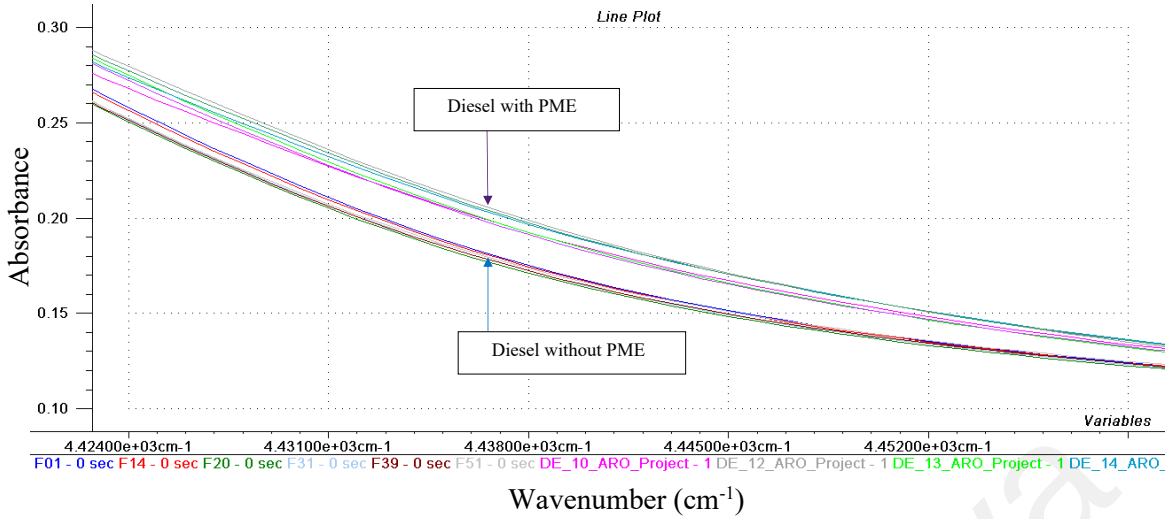


Figure 4.12: Spectra of diesel with and without PME blends between 4452.0 cm^{-1} to 4424.0 cm^{-1} (C=O group) combination overtone region.

PCA was constructed to explore and provide more insight and information on the classification between diesel with and without PME blend. In this study, further analysis was done by applying the chemometrics multivariate methodology using the spectra information measured by FT-NIR spectroscopy. PCA was performed on forty (40) Diesel without PME blend and thirty-six (36) diesel with PME blend from various sources. Full cross-validation was used for PCA construction.

Figure 4.13 illustrates the PCA outcome at the total of 1 PCs and total of 2 PCs score plots. The first two PCs (total of 1 PC and total of 2 PCs) have a total *X* variance explained 90% and demonstrated significant differentiation and clusters between diesel with and without PME blend.

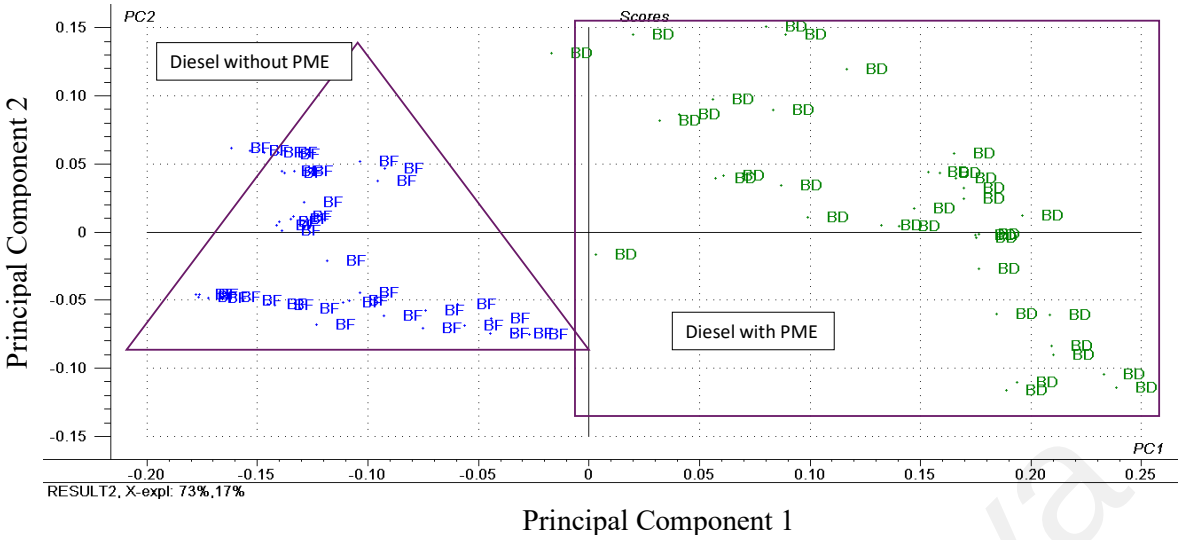


Figure 4.13: Score plot at PC1 and PC2 derived from Principal Component Analysis.

In addition to the above clusters, more information was gathered from the individual cluster, i.e., diesel with and without PME blend. For diesel with PME blend, two clusters are observed at PC2 (positive and negative score values) due to the variation of the PME concentration blended with diesel which can vary from 7 vol/vol % up to 20 vol/vol% depending on the oil producers. For diesel without PME blend, two clusters are observed at PC2 (positive and negative score values), indicating a variation of the diesel compositions. The possibility of the composition variation due to different process unit sources of the diesel, i.e., straight run down from CDU and distillate hydrotreater.

However, as illustrated in Figure 4.14, both samples exhibit overlapping clusters at PC3 and PC4, with the total *X* variance explained by another 8%. This result demonstrates that both samples, which were clustered, mainly contain similar chemical compositions. At PC1, PC2 and PC3, the total explained variance for *X* variables reported more than 95%, as shown in Figure 4.15.

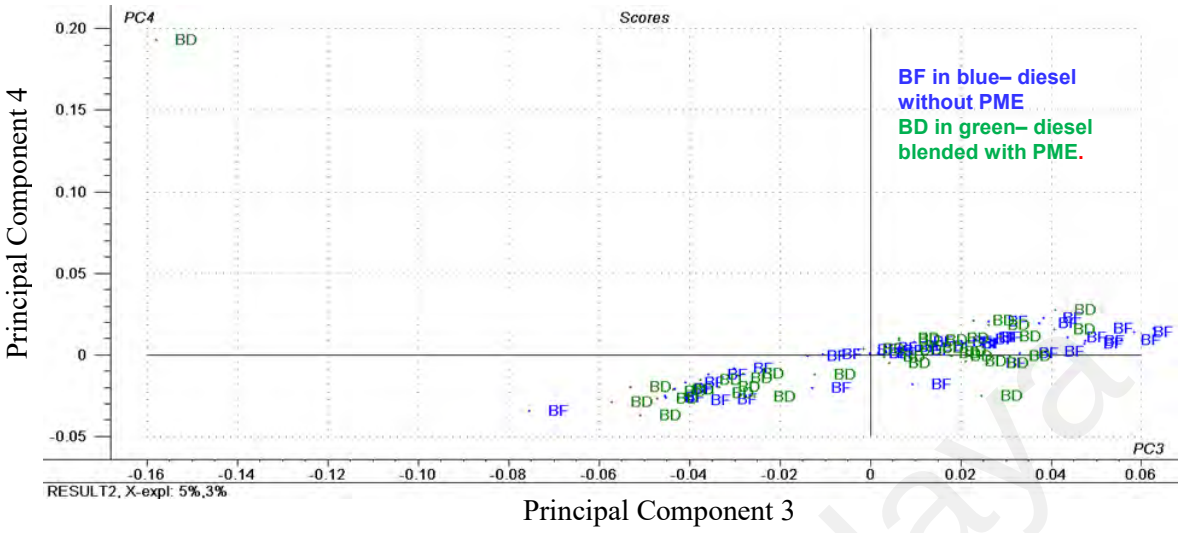


Figure 4.14: Score plot at PC3 and total PC4 derived from Principal Component Analysis.

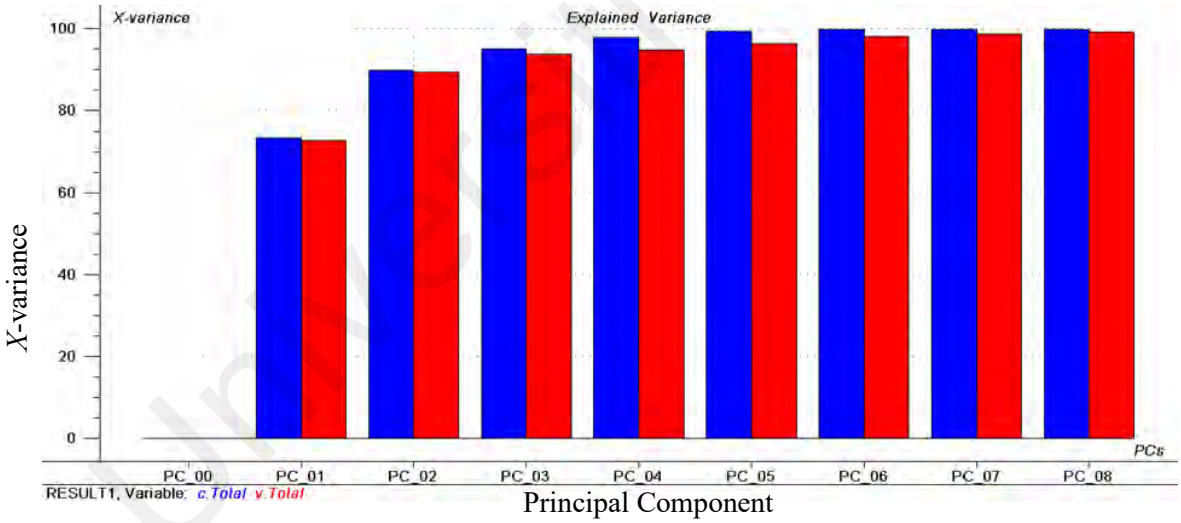


Figure 4.15: Explained variance for *X* variables plot

The outcomes of the PCA results show that the differentiation of the classes is so clearly differentiated. Hence this method can be used as a rapid qualitative measurement to identify the sources and types of diesel fuels without testing using the reference laboratory method, which is time-consuming and costly.

4.6 Quantitative Measurement – Calibration and Prediction

4.6.1 Optimal number of principal components in modelling

An important problem for all the data compression methods in PLSR and PCR is selecting the optimal number of variables or components to use.

If too many components are used, too much of the redundancy in the *X* variables are used, and the solution becomes overfitted, which include 'noise' since the 'noise' may be fitted as well (Loh, 2016). The equation will be very data-dependent and will give poor prediction results.

In contrast, using too few components is called underfitting, which the important variability of the model may cause is not large enough to capture (Næs, 2002). These two important phenomena are illustrated in Figure 4.16.

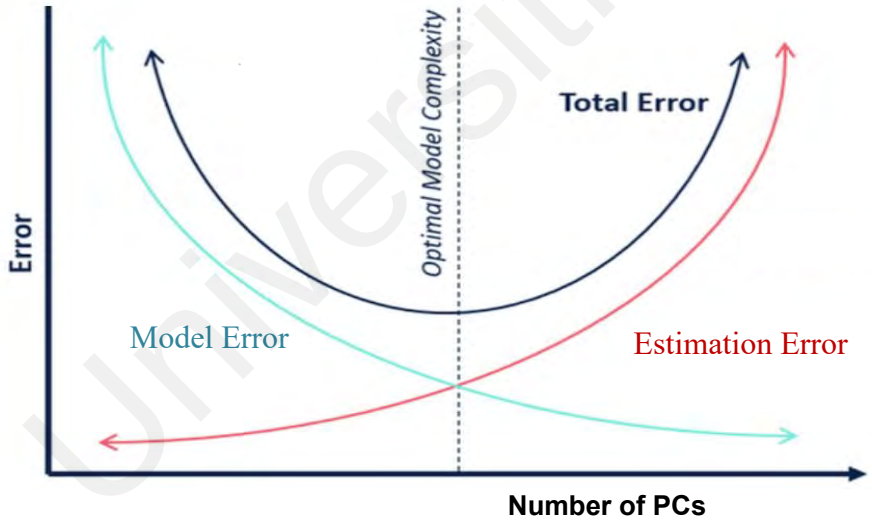


Figure 4.16: Conceptual illustration of model error and estimation error tradeoff in predictive modelling

There are two effects, i.e. estimation error and model error. The estimation error is the error associated with estimating the regression parameter (the statistical uncertainty error) which always increases as more components are added. The calibration model error is comparing the

predicted and measured Y_{cal} values (lab data) which give us an expression of the modelling error because we have only used A components in the model:

$$X_{cal} + Y_{cal} = Model \tag{4.1}$$

$$Then\ we\ feed\ the\ X_{cal} + Model = \hat{Y}_{cal} \tag{4.2}$$

$$Modeling\ error = \hat{Y}_{cal} - Y_{cal} \tag{4.3}$$

This is calculated for each object. Summing the squared differences and taking their mean over all *N* objects gives the calibration residual *Y* variance.

$$Residual\ variance_{cal} = \frac{\sum(\hat{Y}_{cal} - Y_{cal})^2}{N} \tag{4.4}$$

The square root of the above equation gives us RMSEC (Root Mean Square Error of Calibration), the modelling error, expressed in original measuring units.

The sum curve of these two opposing trends will therefore generally display a more or less well-defined minimum which corresponds to the optimal number of components.

As the number of variables or components increases, the model's ability to capture X-variability increases, resulting in decreased model error. However, the estimation error increases due to the higher number of parameters that need to be estimated. The optimal number of variables or components lies between these two extremes.

Selecting the optimal number of variables or components is a crucial problem for all data compression methods. If too many components are used, the model may become overfitted as it utilizes too much redundancy in the X variables, resulting in poor prediction results that are

highly dependent on the data. Using too few components, on the other hand, called underfitting means that the model is not large enough to capture the important variability in the data.

The overfitting effect is strongly dependent on the number of samples used (Martens & Naes, 1984). The more samples used, the more precise the parameters estimates, and thus, the prediction is. Therefore, the more samples used, the less important the overfitting effects are.

The optimal number of PCs or LVs can be appropriately selected by studying the validation residual variance plot. One should choose the number of PCs corresponding to the first clear V-minimum or a break from monotonically decreasing residual variance, i.e. where the prediction error is minimised. The Unscrambler software program suggests an optimal number of PCs based on the calculation to the following formula (Esbensen, 2002):

$$PLS\ or\ PCR: Min[Vytot_{valPC0} \times 0.01 \times a + Vytot_{valPCa}] \quad (4.5)$$

$$PCA\ \ : Min[Vxtot_{valPC0} \times 0.01 \times a + Vxtot_{valPCa}] \quad (4.6)$$

Where:

a = current dimesionality (PC number)

Vytot = Total residual Y-variance at validation

*Vxto*t = Total residual *X*-variance at validation

Index *PC0* = at PC number zero

Index *PCa* = at PC number a

In conclusion, choosing fewer PCs gives a more robust model, which is less sensitive to noise and errors, especially the unavoidable sampling errors.

Therefore, the approach taken to choose the appropriate number of principal components (PCs) or latent variables (LVs) involves selecting the model that includes the minimum number of PCs/LVs, leading to a negligible difference between the root mean square error of calibration (RMSEC) and root mean square error of prediction (RMSEP).

The RMSEC and RMSEP can be referred to as the average modelling and predictive error (Hadad et al., 2008, Loh, 2016). The following equation gives the ratio RMSEC or RMSEP:

$$\text{RMSEC or RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (4.7)$$

Where $\hat{y}_i$ and y_i are the estimated NIR result, and the laboratory reference result of the ith sample for each parameter model and n is referred to the number of samples in the calibration/prediction set.

4.6.1.1 Calibration of boiling point at 95 % recovery

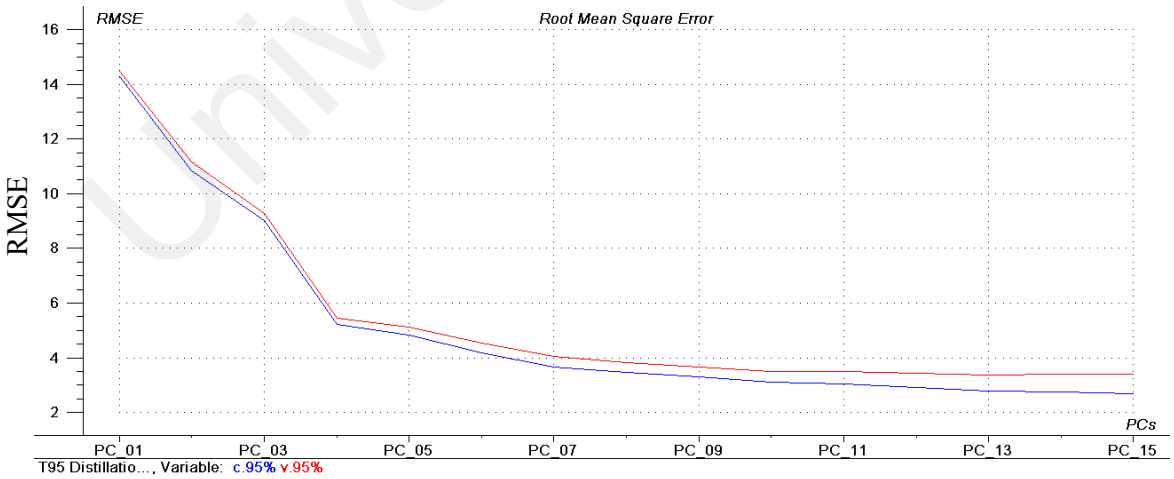
(a) MSC-PLSR (Total of 7 PCs)

For the MSC-PLSR model, the RMSE plot indicates two possible total PCs, which are 4 PCs and 7 PCs. The total 7 PCs indicates the optimum PC since no noise has been shown in the x-loading and regression coefficient plots (Figure 4.18). Total of 4 PCs explained about 92% of *Y* variance, whereas total of 7 PCs explained about 96% of *Y* variance. Hence, total 4 PCs might be under-fitting and cause a high prediction error (Figure 4.17), although there is no noise indicated in both *X*-loading and regression coefficient plots (Figure 4.21 and Figure 4.22).

X-loading is loadings plot for the *X*-variables (NIR spectra) for a specified component versus the variable number which is wavenumber or wavelength. The importance and usefulness of the plots for detecting important variables. The plot illustrates the correlation between the designated component and the various *X* variables. A high positive or negative loading of a

variable indicates its significance for the respective component. For instance, a sample with a high score value for the component will have a large positive value for a variable with a large positive loading. The variables with high loadings in the initial components are the ones that exhibit maximum variation. Therefore, these variables are accountable for the most significant differences between the samples. If the PC components increased the *X*-loading plot line will show "noise" in which the correlation between *X* variables no longer provides information. *X*-loading plots have a strong correlation with the scores plot not with the regression coefficient plot.

Regression coefficient plots are the numerical coefficients of the model equation that express the link between variation in the predictors (*X*-variables i.e. NIR spectra wavelength or wavenumber) and variation in the response (*Y*-variables i.e. laboratory reference data). The response value from the *X*-measurements is calculated using the regression coefficients, which provide information on the variables that have a significant influence on the response variables based on their magnitudes. The regression coefficient plot line will demonstrate "noise" when the variation in the *X* variables and variation in the *Y* variables are no longer well correlated.



Principal Component

Figure 4.17: RMSE versus PCs plot for MSC-PLSR, boiling point at 95% recovery

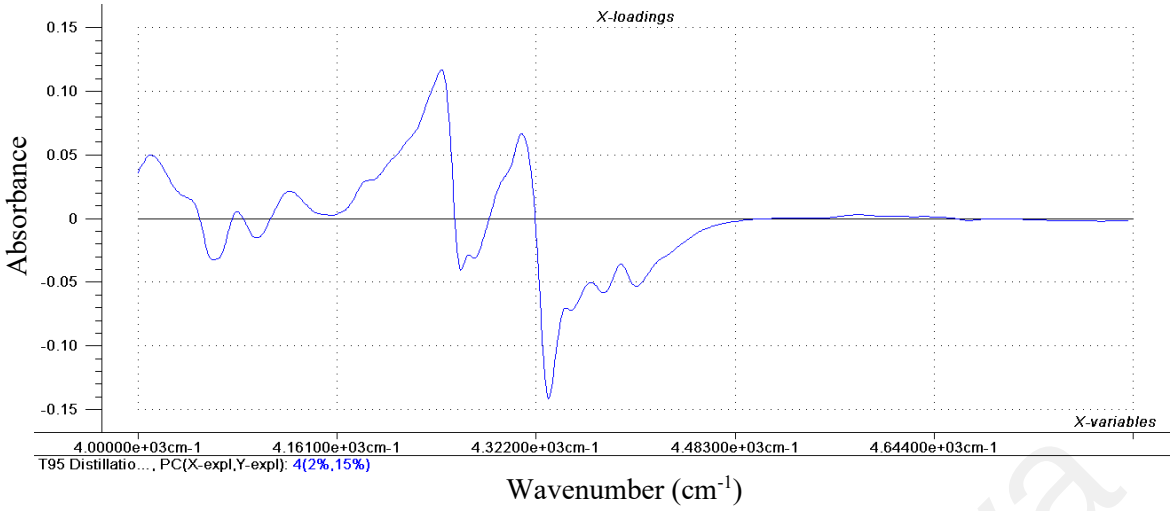


Figure 4.18: *X*-loading plot at total 4 PCs for MSC-PLSR, boiling point at 95% recovery

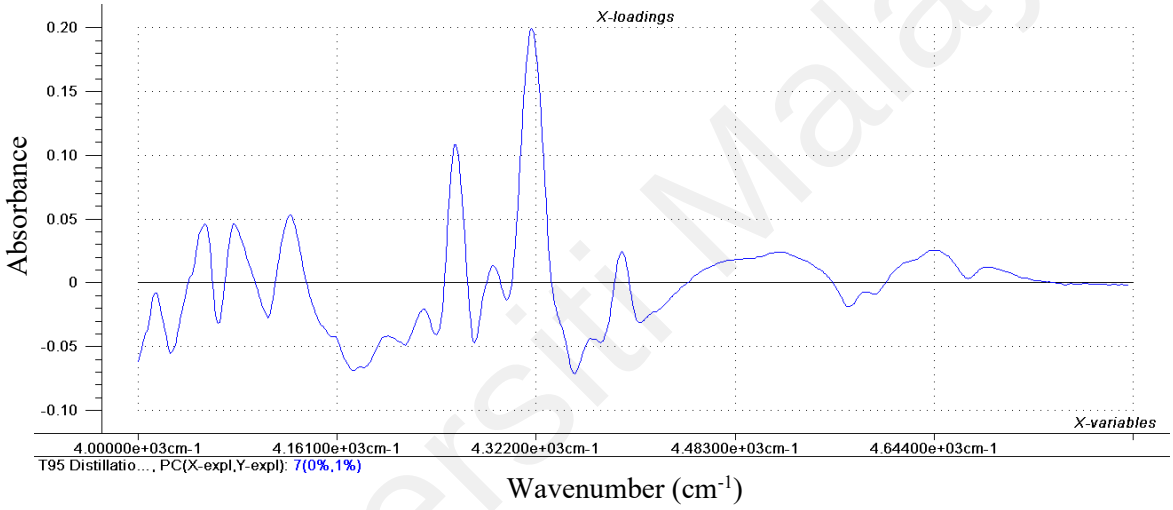


Figure 4.19: *X*-loading plot at total 7 PCs for MSC-PLSR, boiling point at 95% recovery

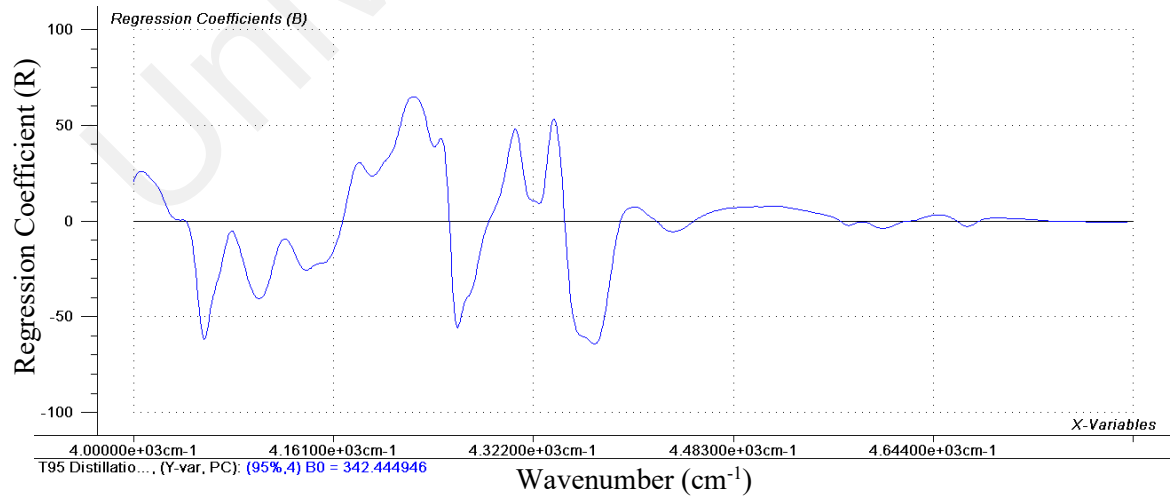


Figure 4.20: Regression coefficient plot at total 4 PCs for MSC-PLSR, boiling point at 95% recovery

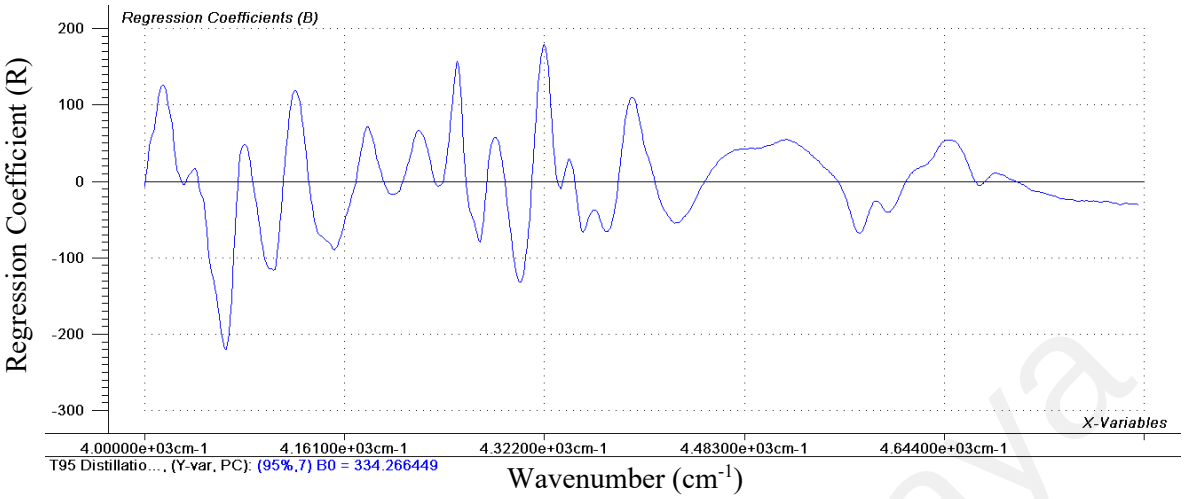


Figure 4.21: Regression coefficient plot at total 7 PCs for MSC-PLSR, boiling point at 95% recovery

(b) MSC-PCR (Total of 4 PCs)

For the MSC-PCR model, the RMSE plot indicates two possible PCs, which are total of 4 PCs and total of 7 PCs. The software diagnostic tool identified total of 7 PCs as the optimum PC rather than total of 4 PCs. Further evaluation on the x-loading and regression coefficient plots indicated slight noise at total of 7 PCs, which might lead to invalid prediction although it had explained about 95% of *Y* variance. (Figure 4.23 and 4.24).

Hence, total of 4 PCs has selected as an optimum PC with *Y* variance was explained about 92%. No noise was shown in x-loading and regression coefficient plots (Figure 4.25 and Figure 4.26).

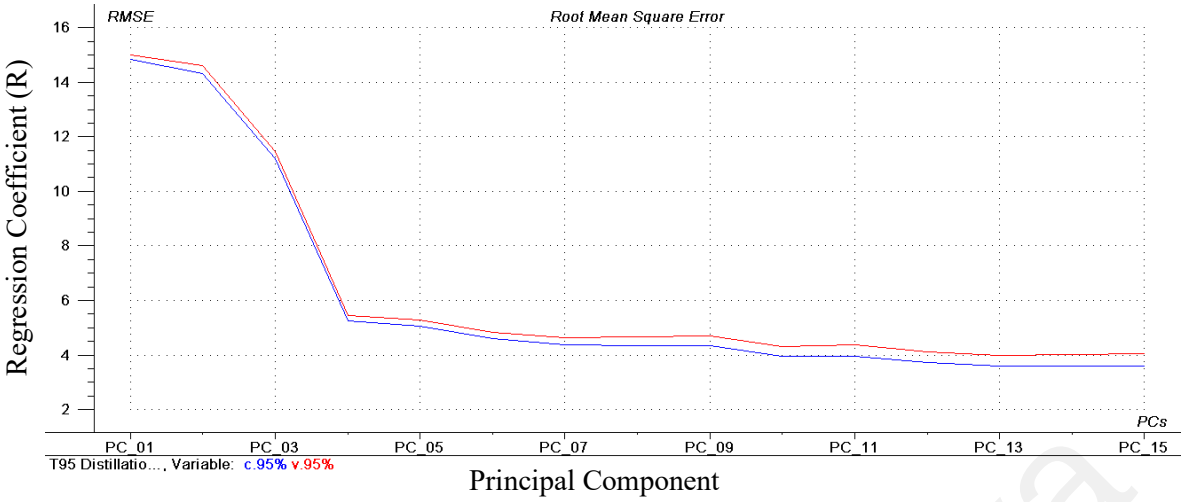


Figure 4.22: RMSE versus PCs plot for MSC-PCR, boiling point at 95% recovery

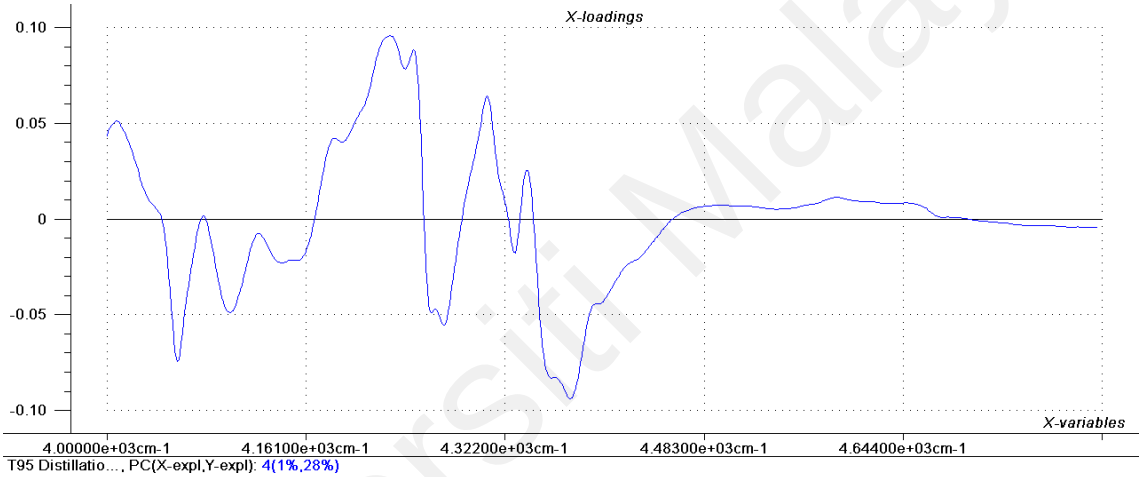


Figure 4.23: *X*-loading plot at total 4 PCs for MSC-PCR, boiling point at 95% recovery

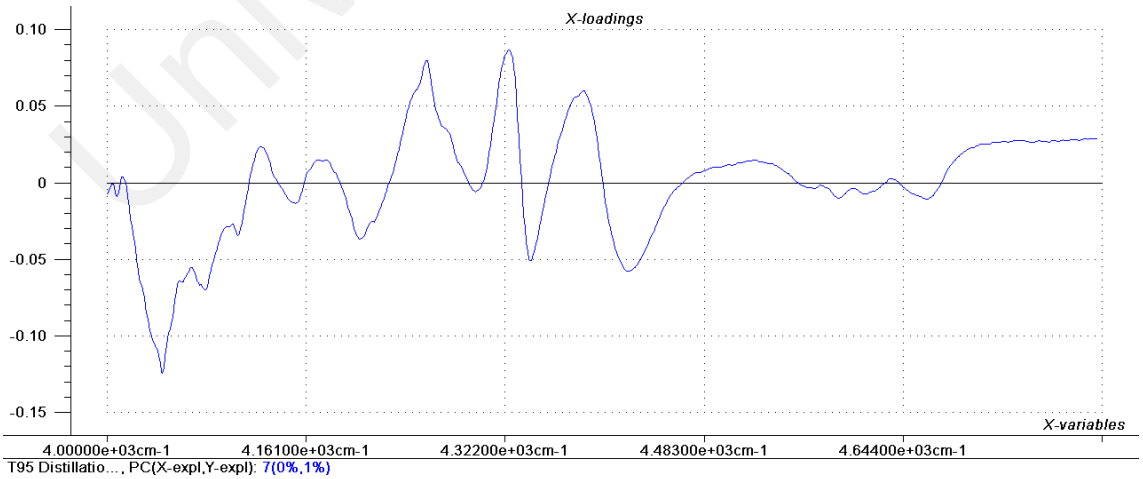


Figure 4.24: *X*-loading plot at total 7 PCs for MSC-PCR, boiling point at 95% recovery

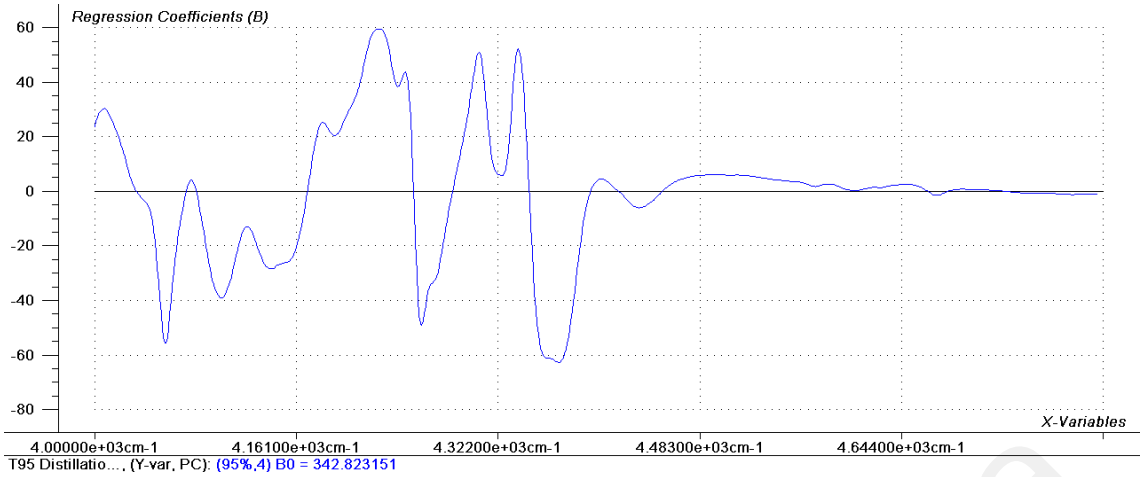


Figure 4.25: : Regression coefficient plot at total 4 PCs for MSC-PCR, boiling point at 95% recovery

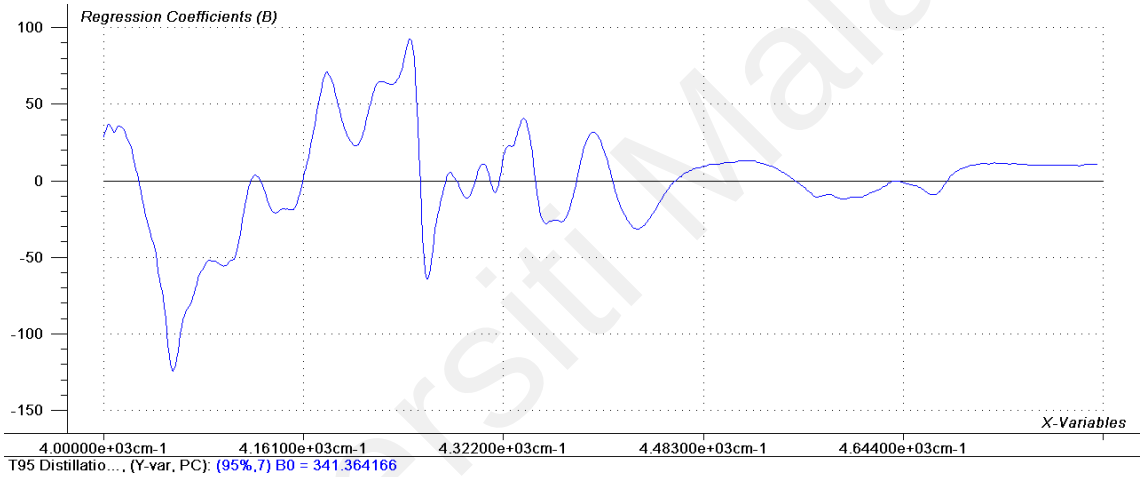


Figure 4.26: Regression coefficient plot at total 7 PCs for MSC-PCR, boiling point at 95% recovery

(c) SGSD-PLSR (Total of 4 PCs)

For the SGSD-PLSR model, the RMSE plot indicates two possible PCs, which are total of 4 PCs and 7 PCs. The software diagnostic tool identified total of 7 PCs as the optimum PC rather than total of 4 PCs, where the RMSE value is much lower (Figure 4.27). However, at total of 7 PCs, the difference of RMSE for calibration and validation indicates a significant difference. Further evaluation on the x-loading and regression coefficient plots indicated much noise at total of 7 PCs, which might lead to invalid prediction, although it had explained about 97% of *Y* variance. (Figure 4.28 and 4.29).

Hence, total of 4 PCs has selected as an optimum PC with *Y* variance was explained about 93%. There is no noise in x-loading and regression coefficient plots (Figure 4.30 and Figure 4.31).

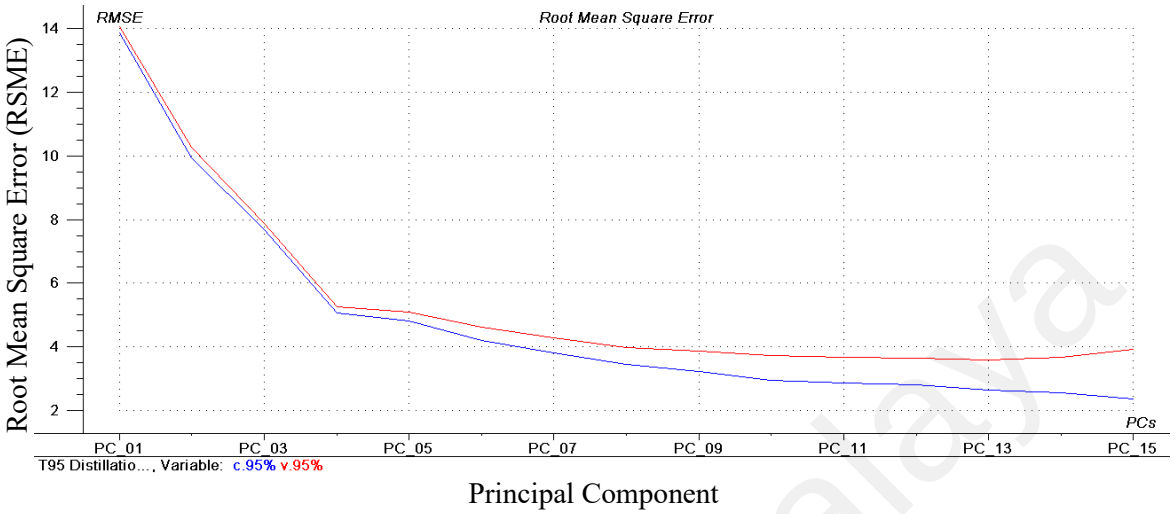


Figure 4.27: RMSE versus PCs plot for SGSD-PLSR, boiling point at 95% recovery

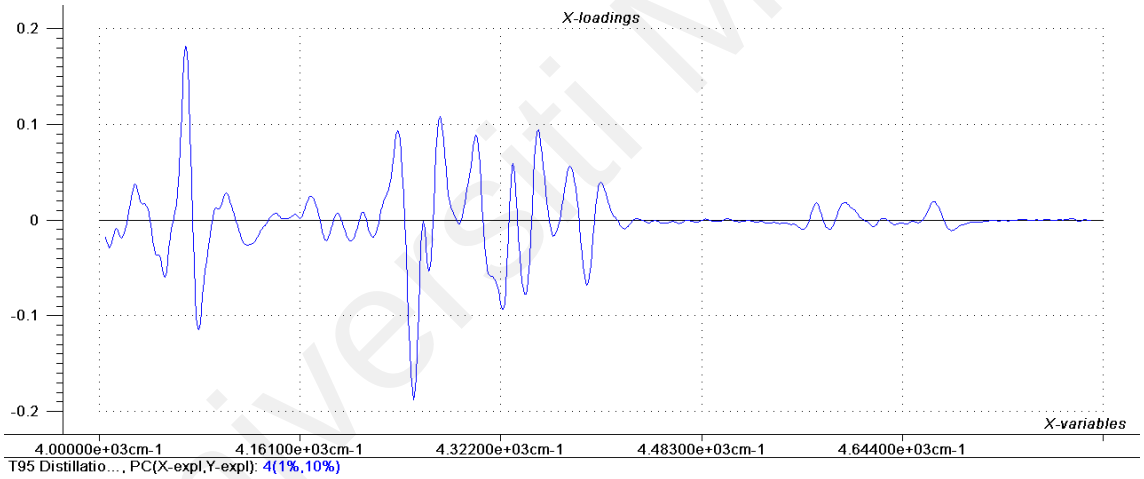


Figure 4.28: *X*-loading plot at total 4 PCs for SGSD-PLSR, boiling point at 95% recovery

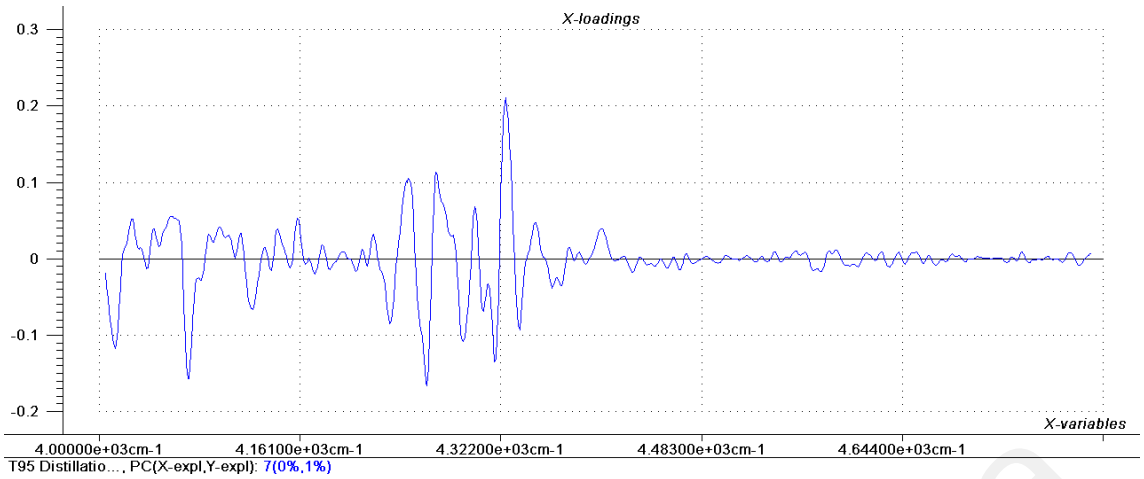


Figure 4.29: *X*-loading plot at total 7 PCs for SGSD-PLSR, boiling point at 95% recovery

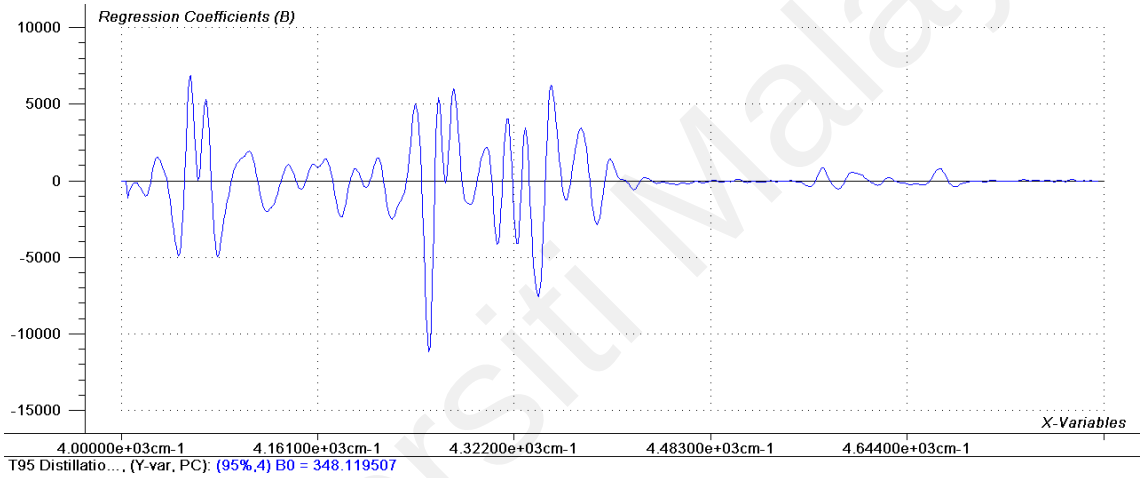


Figure 4.30: Regression coefficient plot at total 4 PCs for SGSD-PLSR, boiling point at 95% recovery

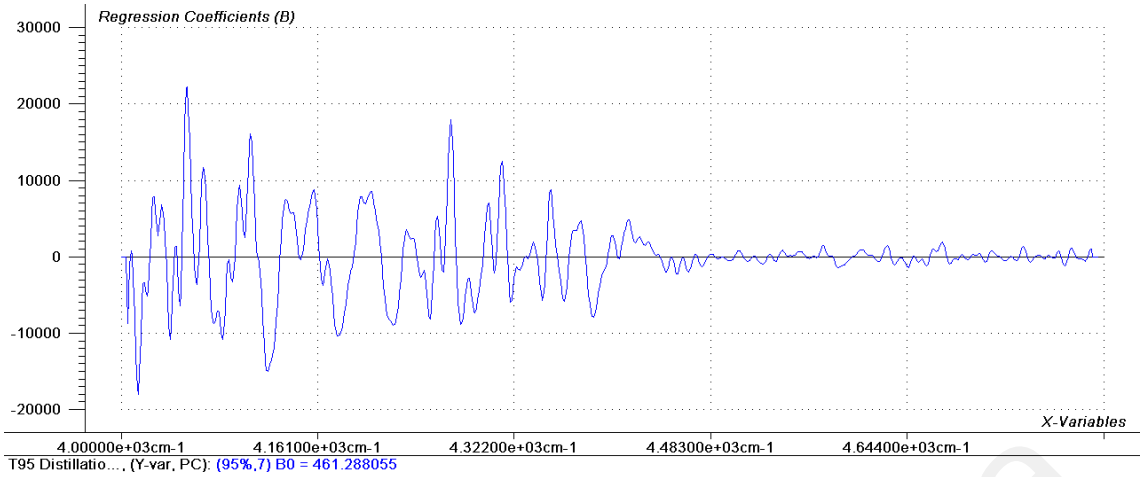


Figure 4.31: Regression coefficient plot at total 7 PCs for SGSD-PLSR, boiling point at 95% recovery

(d) SGSD-PCR (Total of 5 PCs)

For the SGSD-PCR model, the RMSE plot indicates two possible PCs, which are total of 4 PCs and total of 5 PCs. The software diagnostic tool identified total of 5 PCs as the optimum PC rather than total of 4 PCs (first minimum curve), where the RMSE value is much higher (Figure 4.32). At total of 5 PCs, the difference of RMSE for calibration and validation indicates no significant difference. Further evaluation of the *X*-loading and regression coefficient plots indicated no noise at total of 5 PCs with the *Y* variance explained 93% (Figure 4.33 and 4.34).

Total of 4 PCs might be under-fitting and cause a high prediction error, although no noise is indicated in both *X*-loading and regression coefficient plots (Figure 4.35 and Figure 4.36).

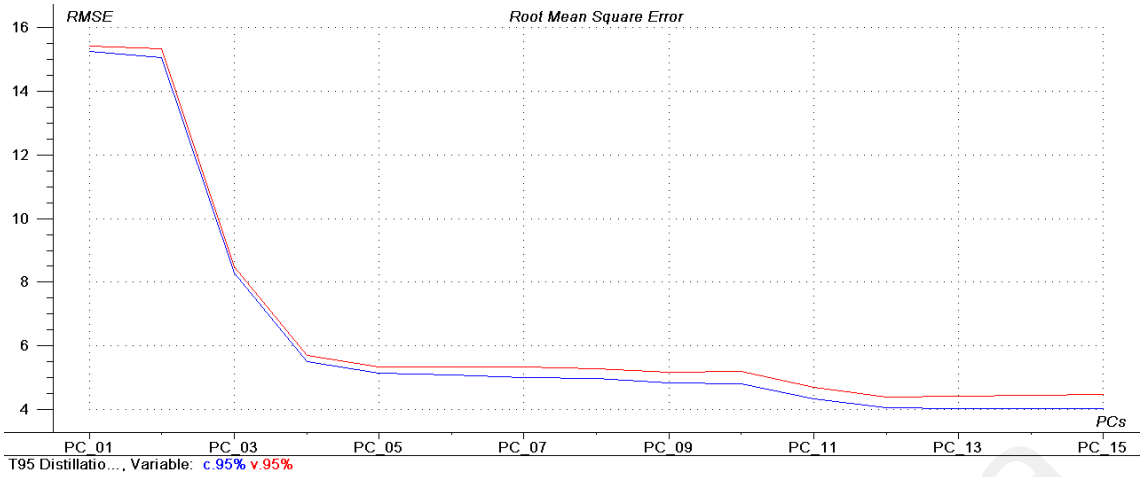


Figure 4.32: RMSE versus PCs plot for SGSD-PCR, boiling point at 95% recovery

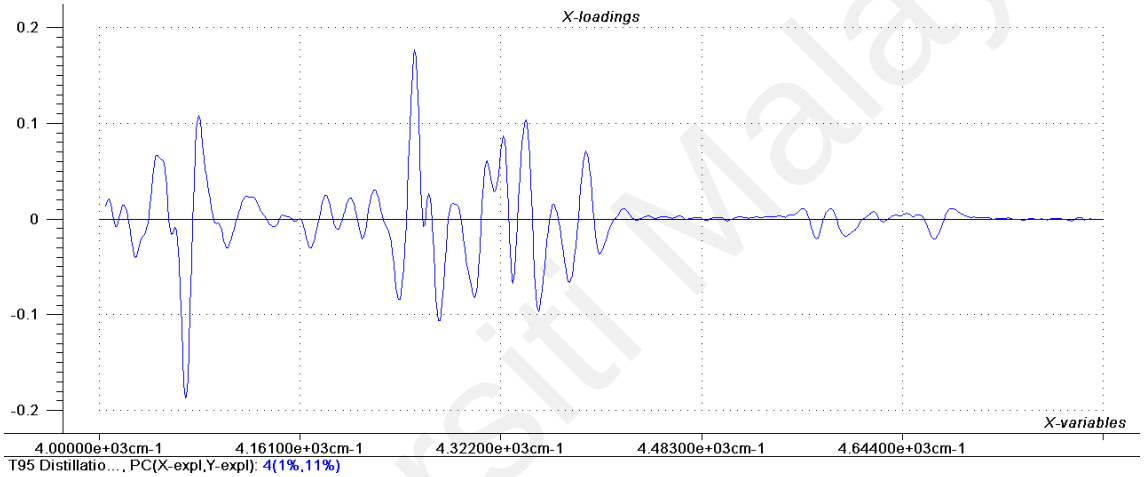


Figure 4.33: *X*-loading plot at total 4 PCs for SGSD-PCR, boiling point at 95% recovery

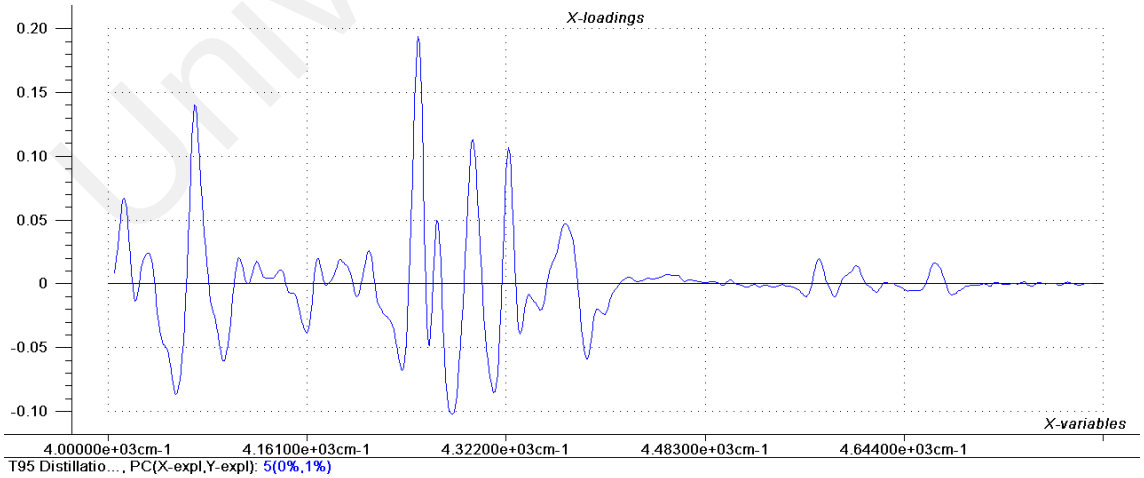


Figure 4.34: *X*-loading plot at total 5 PCs for SGSD-PCR, boiling point at 95% recovery

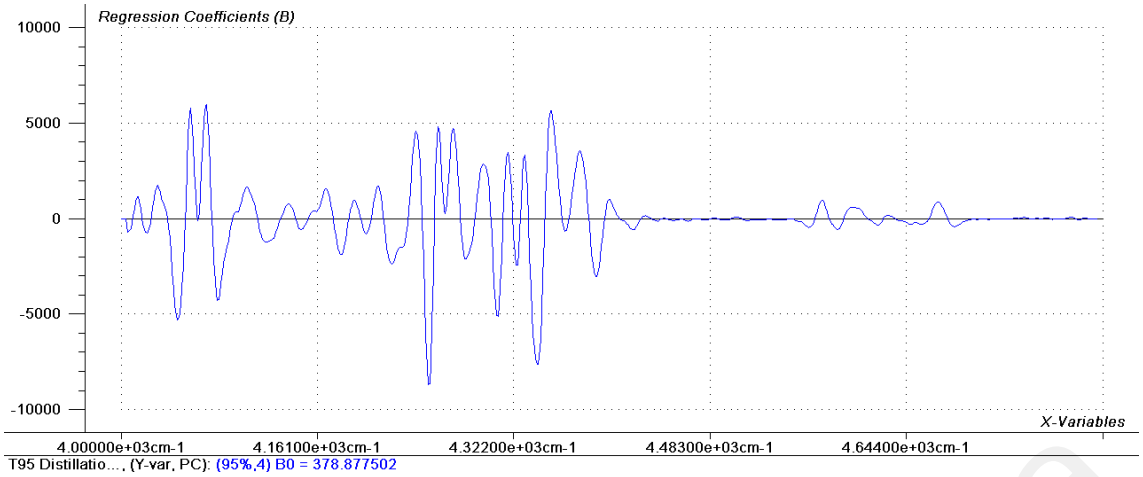


Figure 4.35: Regression coefficient plot at total 4 PCs for SGSD-PCR, boiling point at 95% recovery

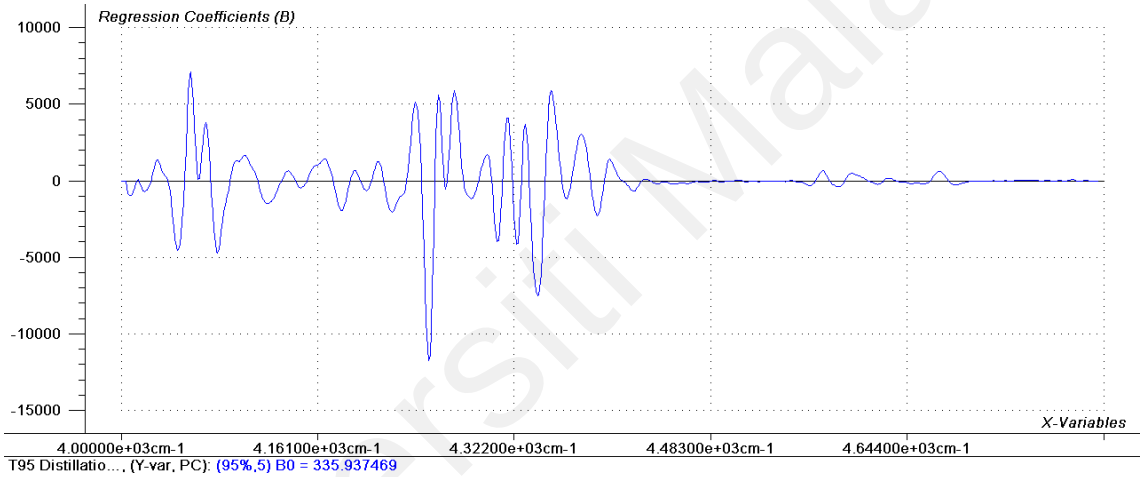


Figure 4.36: Regression coefficient plot at total 5 PCs for SGSD-PCR, boiling point at 95% recovery

4.6.1.2 Calibration of Flash Point

(a) MSC-PLSR (Total of 9 PCs)

For the MSC-PLSR flash point model, the RMSE plot indicates two possible PCs, which are total of 4 PCs and total of 9 PCs. The software diagnostic tool identified total of 9 PCs as the optimum PC rather than total of 4 PCs (first minimum curve), where the RMSE value is much higher (Figure 4.37). At total of 5 PCs, the difference of RMSE for calibration and validation indicates a slight difference but is acceptable. Further evaluation of the *X*-loading and regression coefficient plots indicated no noise at 9 PCs. The *Y* variance explained about

92%, whereas total of total of 4 PCs reported about 80% explained on the *Y* variance (Figure 4.38 and 4.39).

Total of 4 PCs might be under-fitting and cause a high prediction error, although no noise is indicated in both *X*-loading and regression coefficient plots (Figure 4.40 and Figure 4.41).

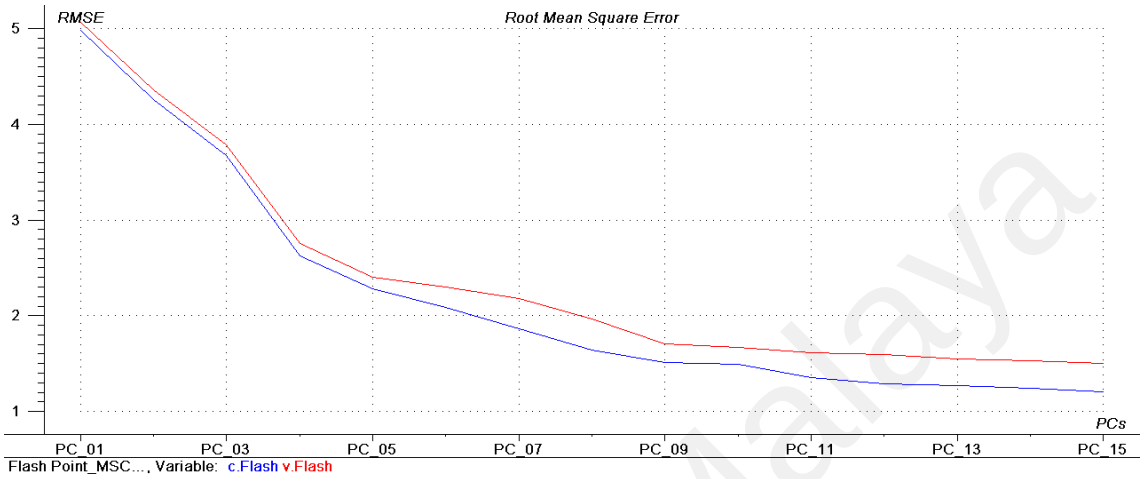


Figure 4.37: RMSE versus PCs plot for MSC-PLSR, flash point

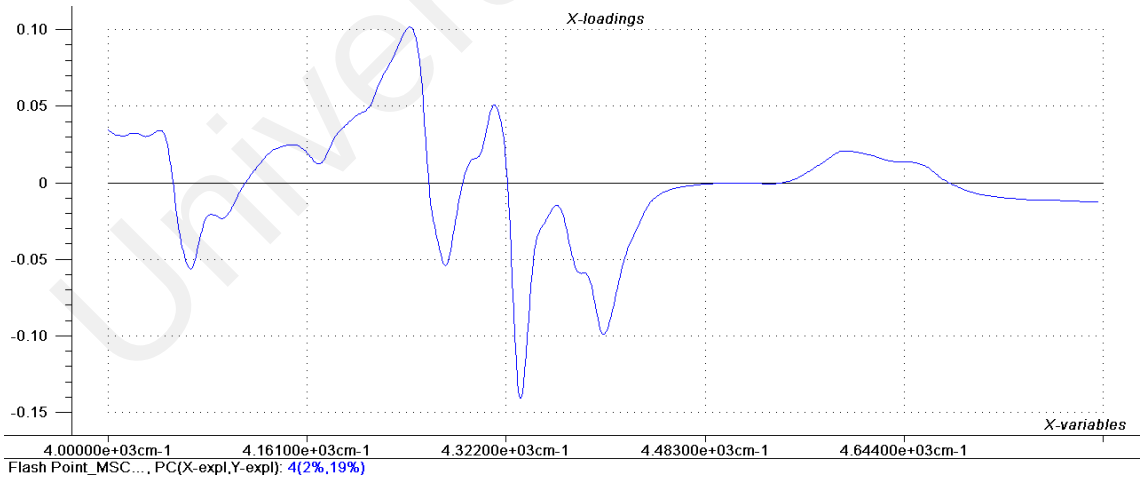


Figure 4.38: *X*-loading plot at total 4 PCs for MSC-PLSR, flash point

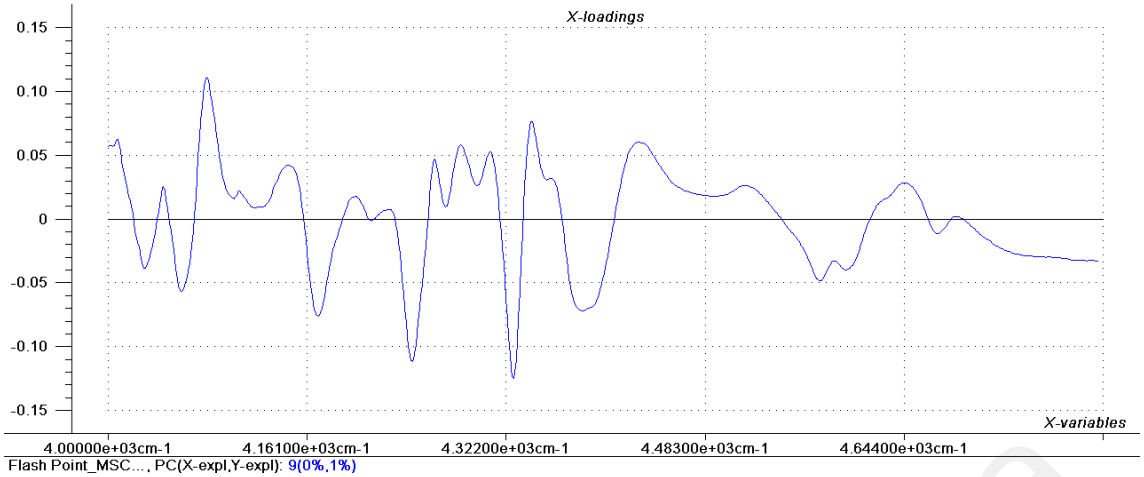


Figure 4.39: *X*-loading plot at total 9 PCs for MSC-PLSR, flash point

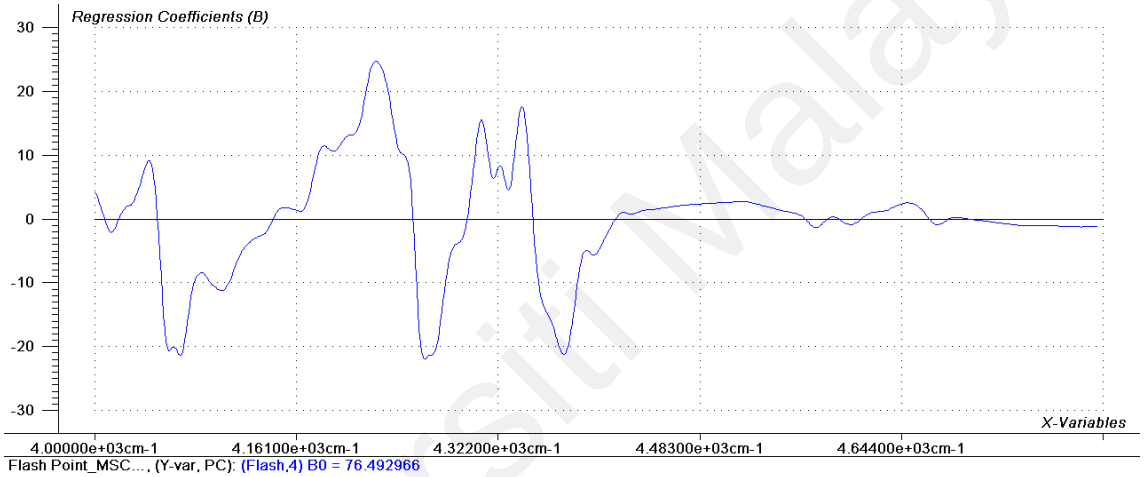


Figure 4.40: Regression coefficient plot at total 4 PCs for MSC-PLSR, flash point

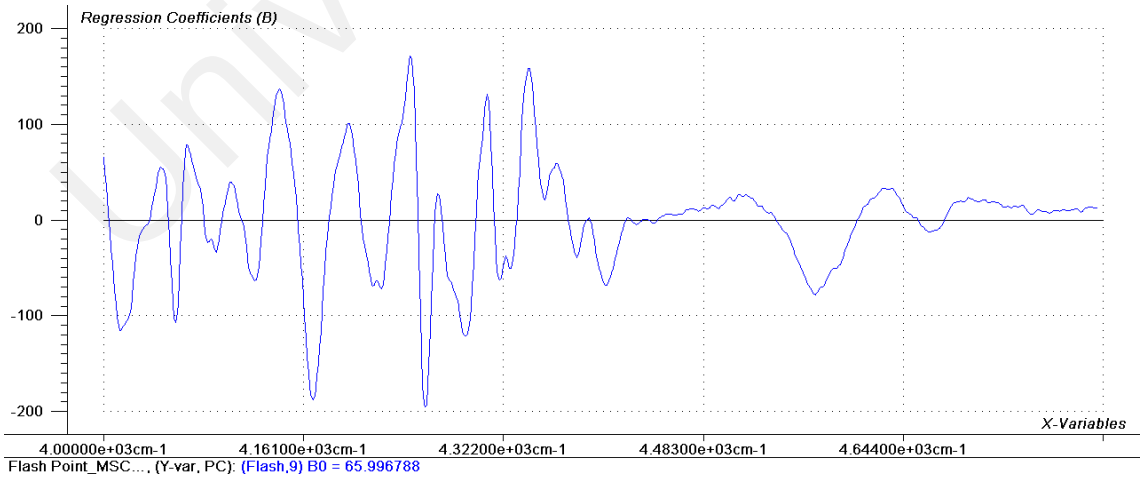


Figure 4.41: Regression coefficient plot at total 9 PCs for MSC-PLSR, flash point

(b) MSC-PCR (Total of 5 PCs)

The RMSE plot indicates a sharped minimum curve with total of 5 PCs as the optimum PC for the MSC-PCR flash point model. The software diagnostic tool also suggested total of 5 PCs as the local minimum as well.

At total of 5 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 5 PCs. The Y variance explained about 86% (Figure 4.42-4.44).

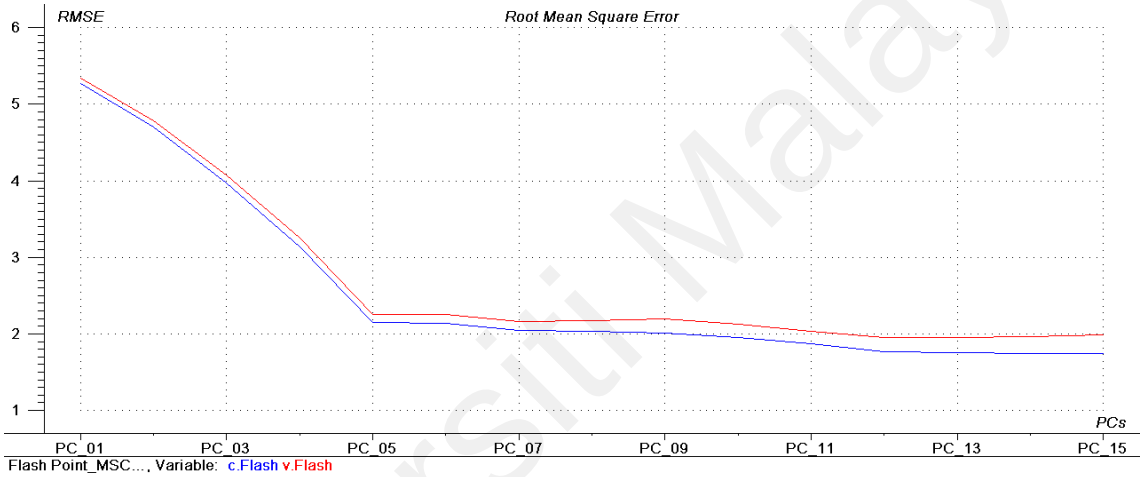


Figure 4.42: RMSE versus PCs plot for MSC-PCR, flash point

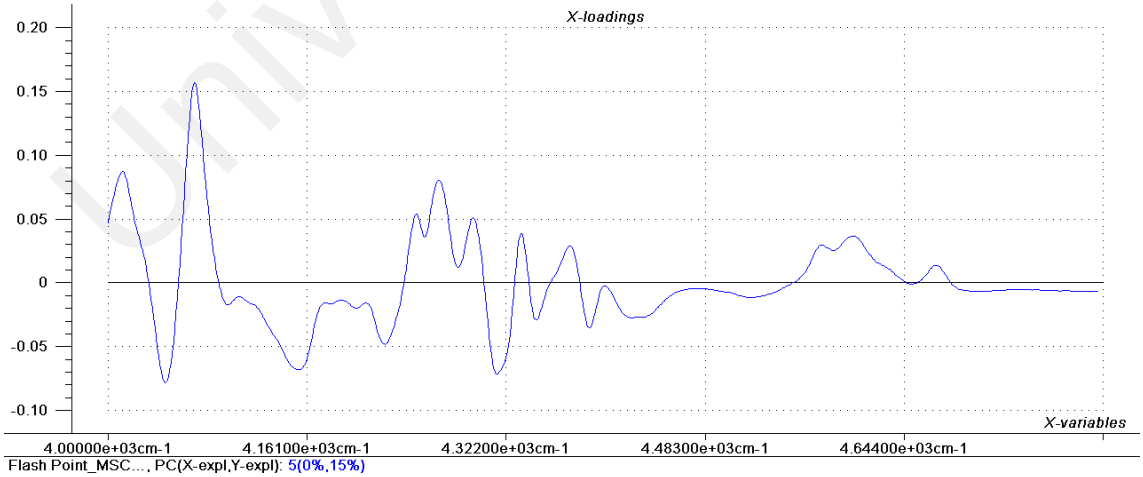


Figure 4.43: *X*-loading plot at total 5 PCs for MSC-PCR, flash point

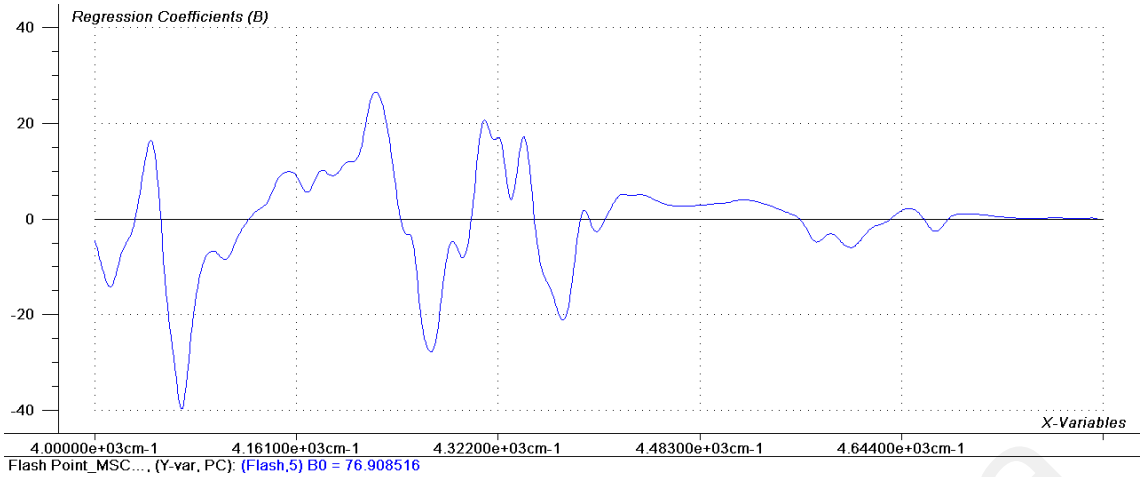


Figure 4.44: Regression coefficient plot at total 5 PCs for MSC-PCR, flash point

(c) SGSD-PLSR (Total of 5 PCs)

The RMSE plot for the SGSD-PLSR flash point model indicates two possible PCs, which are total of 5 PCs and total of 9 PCs. The software diagnostic tool identified total of 9 PCs as the optimum PC rather than total of 5 PCs (first minimum curve), where the RMSE value is much lower (Figure 4.45). At total of 5 PCs, the difference of RMSE for calibration and validation indicates a slight difference but is acceptable compared with total of 9 PCs, which is a significant difference.

Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 5 PCs. The *Y* variance explained about 91% (Figure 4.46 and 4.47).

Total of 9 PCs, although the RMSE value is lower than total of 5 PCs and *Y* variance explained (94%) is higher than total of 5 PCs (91%), much noise is indicated at both x-loading and regression coefficient plots (Figure 4.48 and Figure 4.49). It might lead to instability and inaccuracy of the model because the noise was embedded in the regression for future predictions.

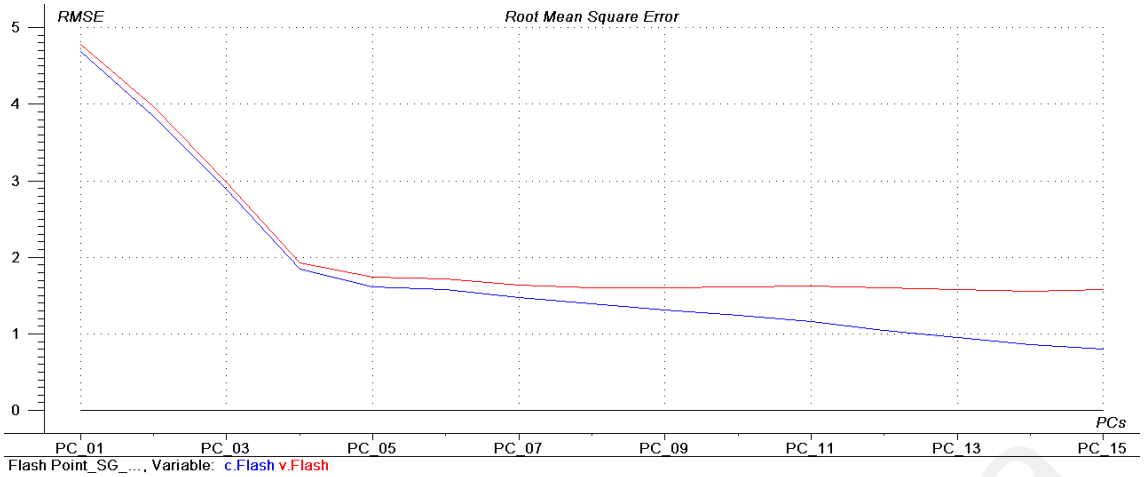


Figure 4.45: RMSE versus PCs plot for SGSD-PLSR, flash point

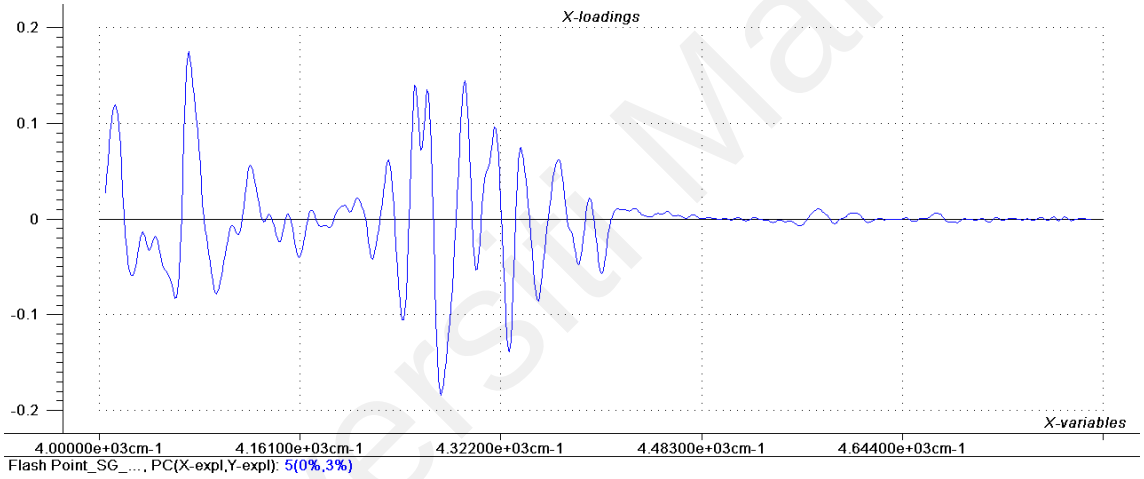


Figure 4.46: *X*-loading plot at total 5 PCs for SGSD-PLSR, flash point

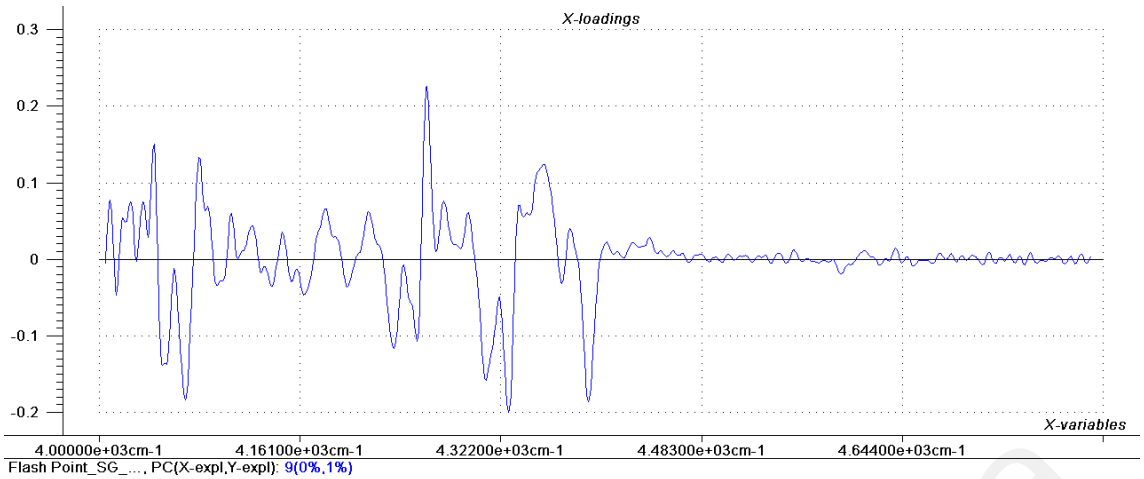


Figure 4.47: *X*-loading plot at total 9 PCs for SGSD-PLSR, flash point

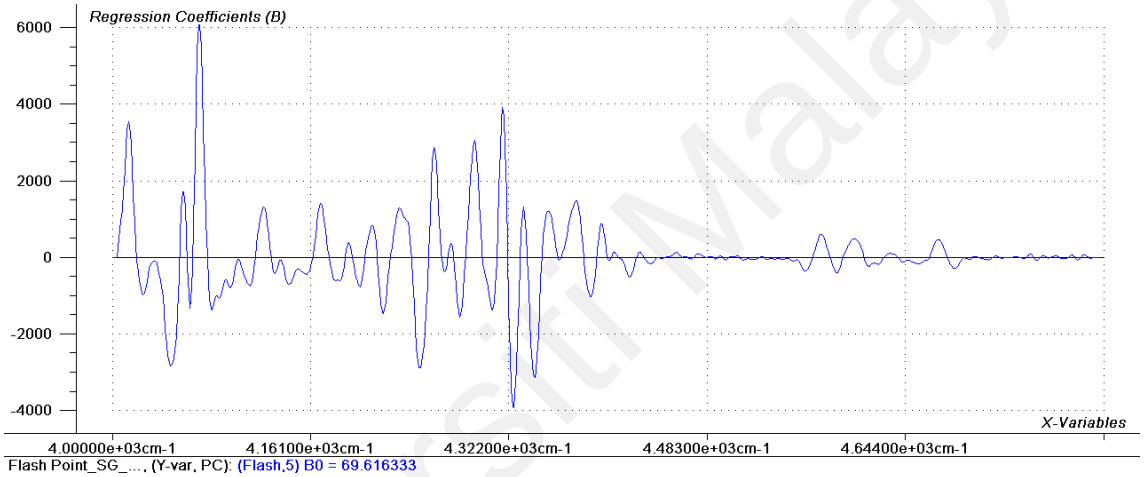


Figure 4.48: Regression coefficient plot at total 5 PCs for SGSD-PLSR, flash point

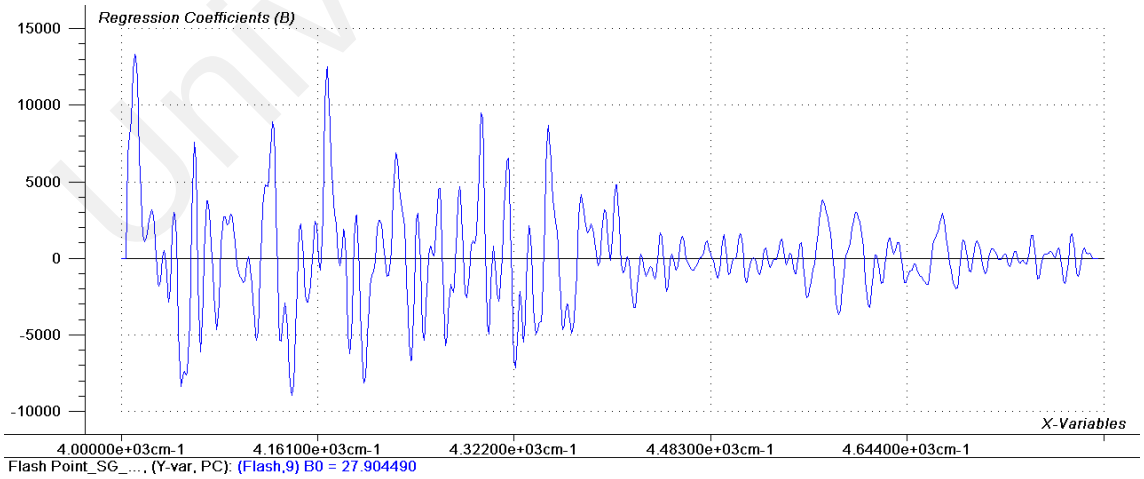


Figure 4.49: Regression coefficient plot at total 9 PCs for SGSD-PLSR, flash point

(d) SGSD-PCR (Total of 4 PCs)

The RMSE plot for the SGSD-PLSR flash point model indicates two possible PCs, which are total of 4 PCs and total of 9 PCs. The software diagnostic tool identified total of 9 PCs as the optimum PC rather than total of 4 PCs (first minimum curve), where the RMSE value is much lower (Figure 4.50). At total of 4 PCs, the difference of RMSE for calibration and validation indicates no significant difference compared with total of 9 PCs, which is significant difference.

Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of PCs. The *Y* variance explained about 87% (Figure 4.51 and 4.52).

Total of 9 PCs, although the RMSE value is lower than total of 5 PCs and *Y* variance explained (91%) is higher than total of 4 PCs (87%), noise is indicated at both x-loading and regression coefficient plots (Figure 4.53 and Figure 4.54). It might lead to instability and inaccuracy of the model because the noise was embedded in the regression for future predictions.

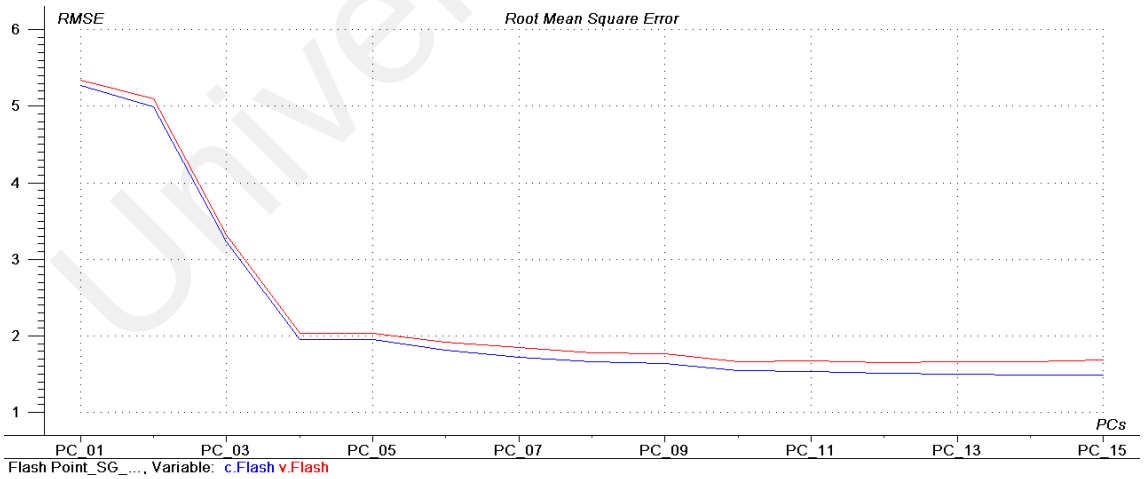


Figure 4.50: RMSE versus PCs plot for SGSD-PCR, flash point

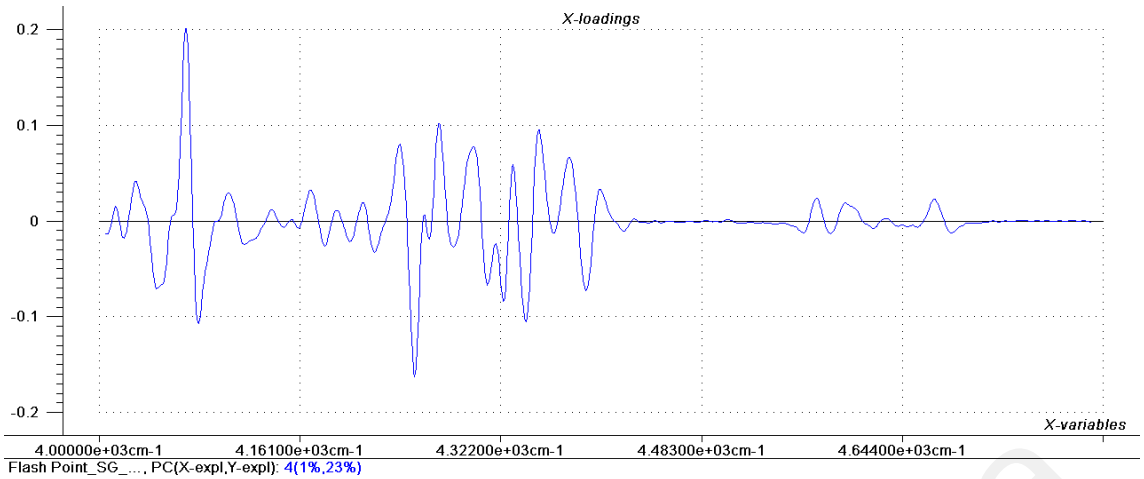


Figure 4.51: *X*-loading plot at total 4 PCs for SGSD-PCR, flash point

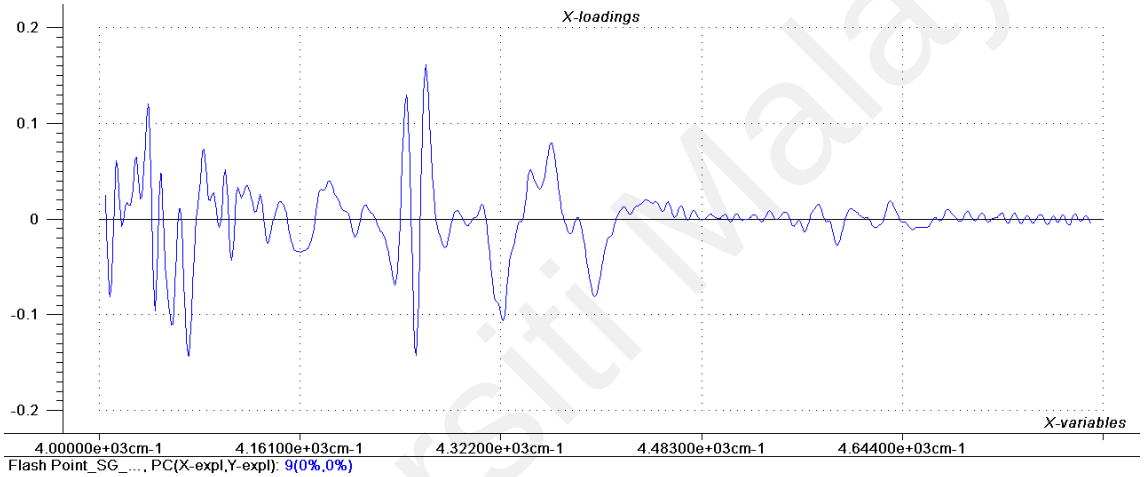


Figure 4.52: *X*-loading plot at total 9 PCs for SGSD-PCR, flash point

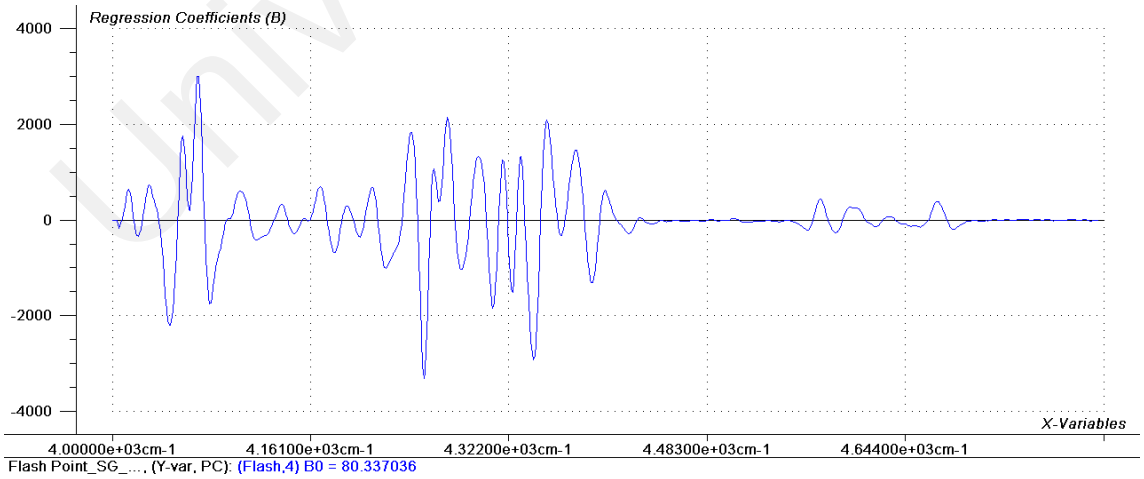


Figure 4.53: Regression coefficient plot at total 4 PCs for SGSD-PCR, flash point

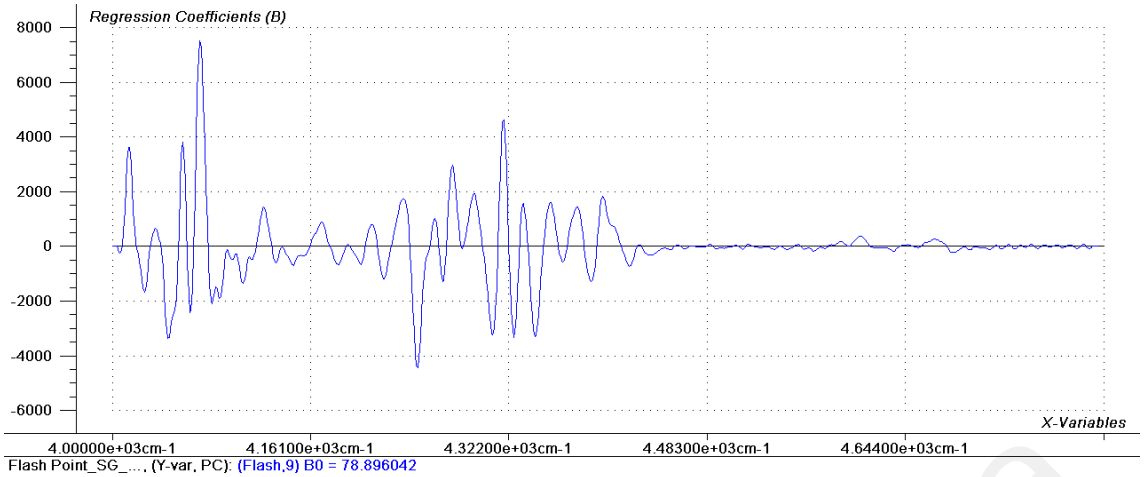


Figure 4.54: Regression coefficient plot at total 9 PCs for SGSD-PCR, flash point

4.6.1.3 Calibration of Cloud Point

(a) MSC-PLSR (Total of 7 PCs)

For the MSC-PLSR cloud point model, the RMSE plot indicates two possible PCs, which are total of 4 PCs and total of 7 PCs. The software diagnostic tool identified total of 7 PCs as the optimum PC rather than total of 4 PCs (first minimum curve), where the RMSE value is much lower (Figure 4.55). At total of 7 PCs, the difference of RMSE for calibration and validation indicates a slight difference but is acceptable. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 7 PCs. The *Y* variance explained about 95%, whereas total of 4 PCs reported about 90% explained on the *Y* variance (Figure 4.56 and 4.57).

Total of 4 PCs might be under-fitting and cause a high prediction error, although no noise is indicated in both *X*-loading and regression coefficient plots (Figure 4.58 and Figure 4.59).

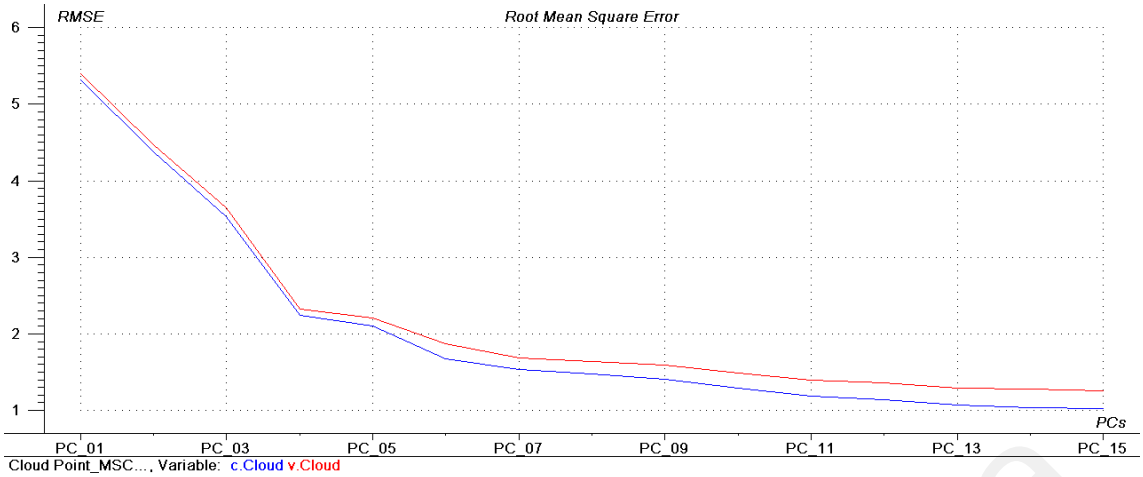


Figure 4.55: RMSE versus PCs plot for MSC-PLSR, cloud point

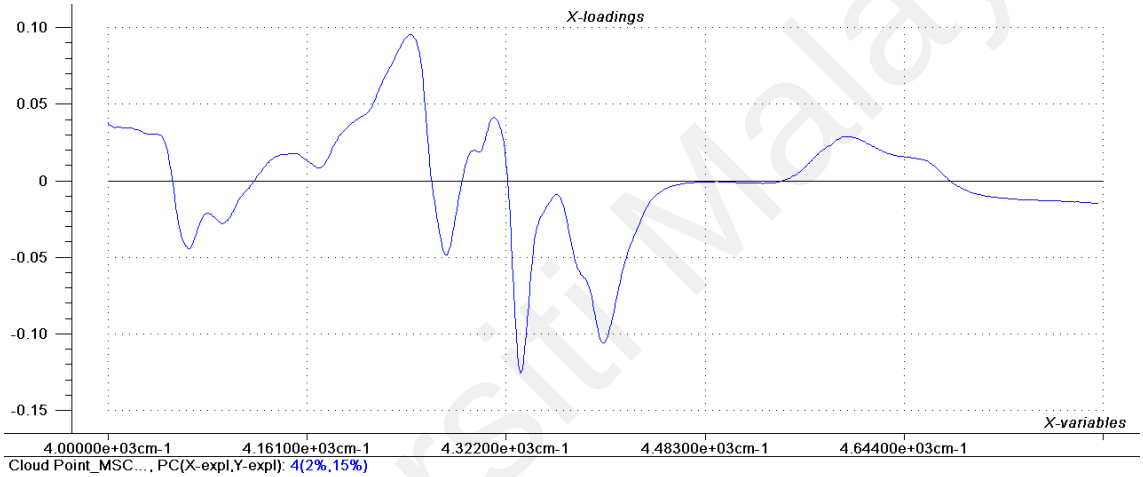


Figure 4.56: *X*-loading plot at total 4 PCs for MSC-PLSR, cloud point

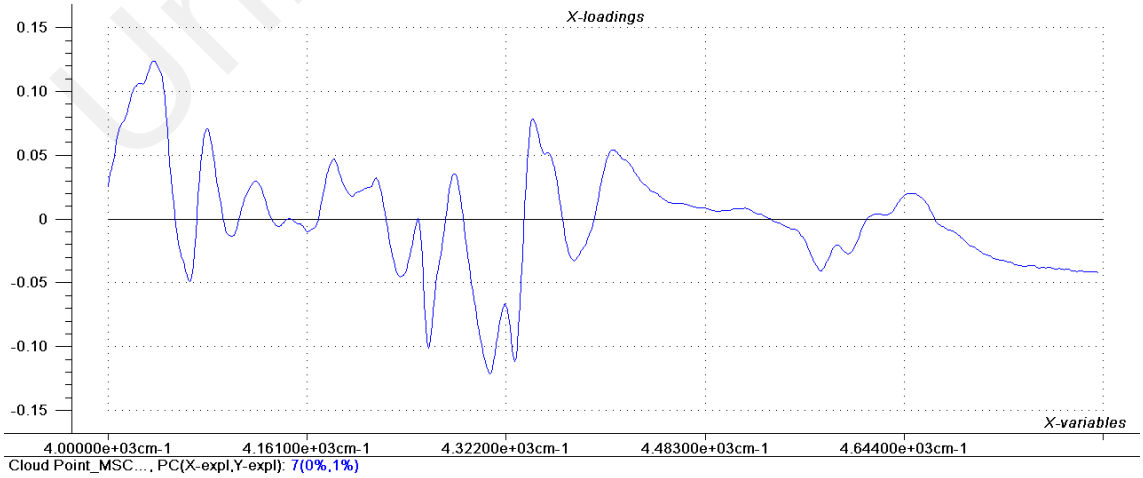


Figure 4.57: *X*-loading plot at total 7 PCs for MSC-PLSR, cloud point

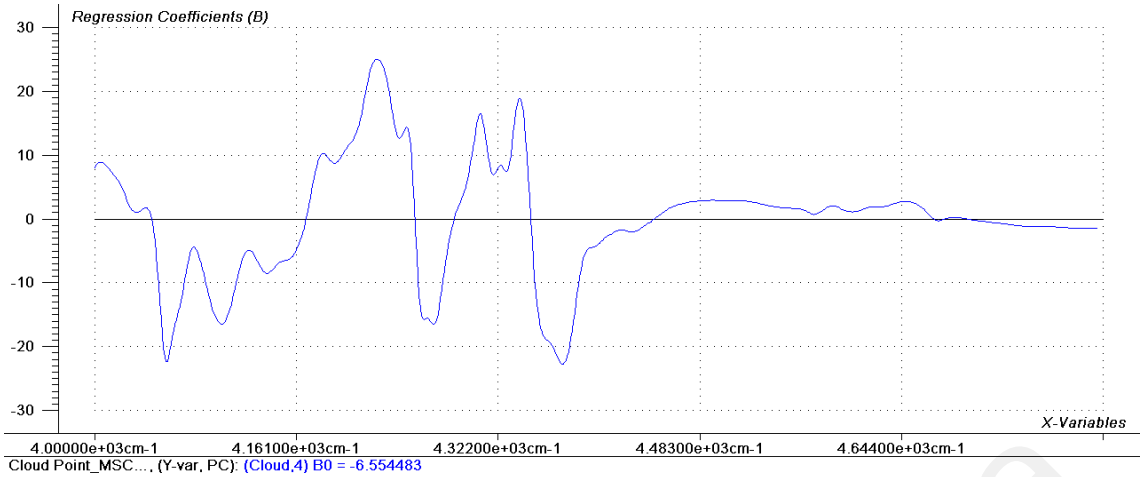


Figure 4.58: Regression coefficient plot at total 4 PCs for MSC-PLSR, cloud point

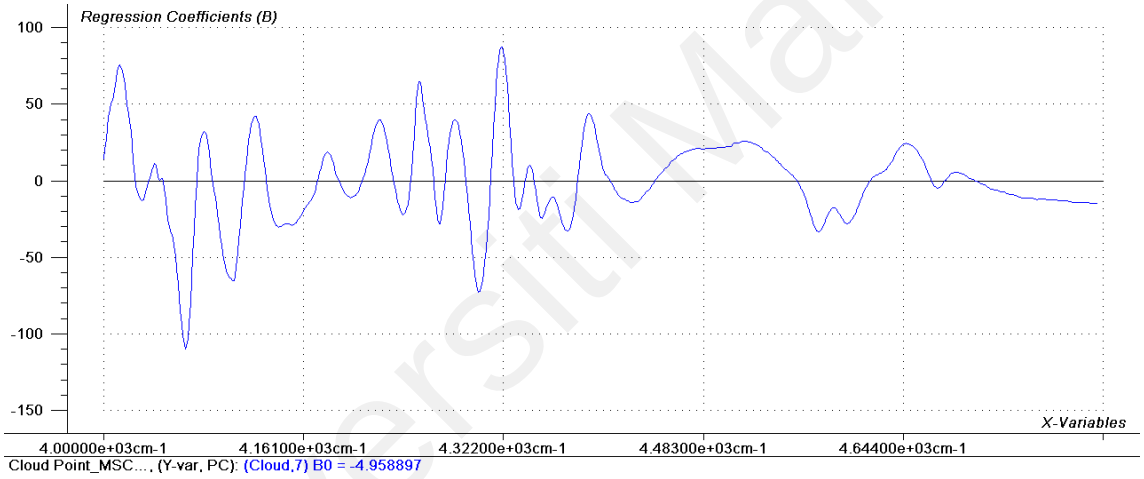


Figure 4.59: Regression coefficient plot at total 7 PCs for MSC-PLSR, cloud point

(b) MSC-PCR (Total of 5 PCs)

The RMSE plot indicates a minimum curve with total of 5 PCs as the optimum PC for the MSC-PCR flash point model. The software diagnostic tool also suggested total of 5 PCs is the local minimum as well.

At total of 5 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 5 PCs. The *Y* variance explained about 90% (Figure 4.61-4.63).

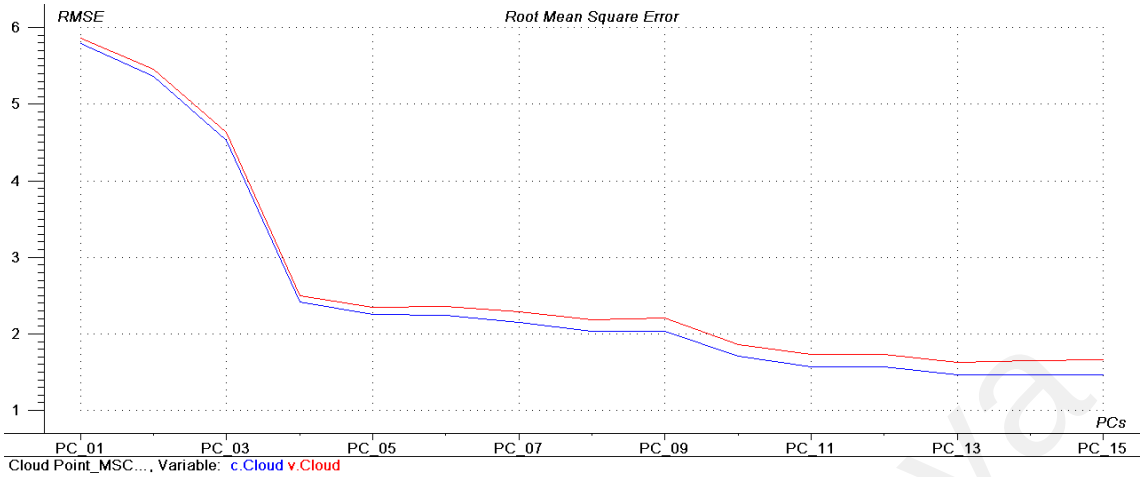


Figure 4.60: RMSE versus PCs plot for MSC-PCR, cloud point

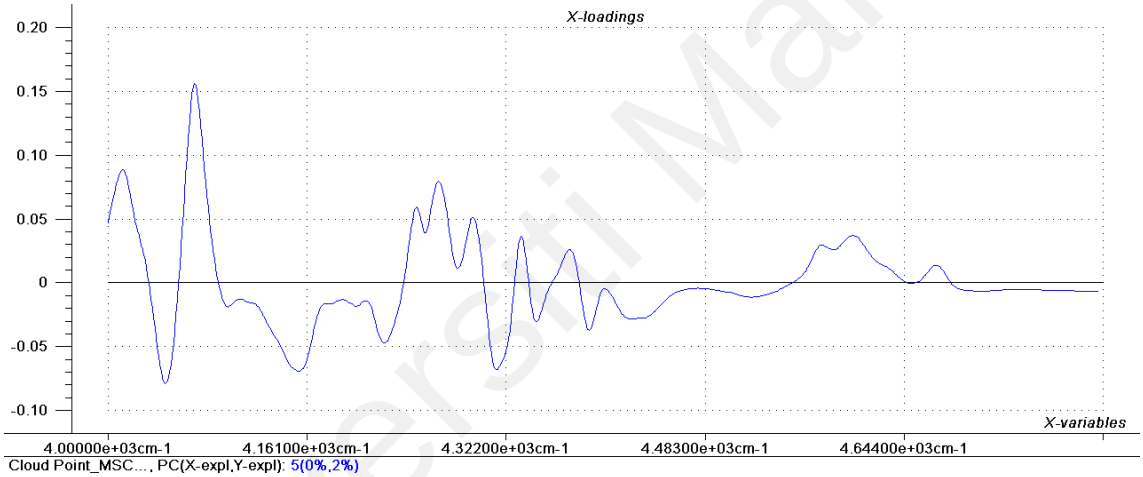


Figure 4.61: *X*-loading plot at total 5 PCs for MSC-PCR, cloud point

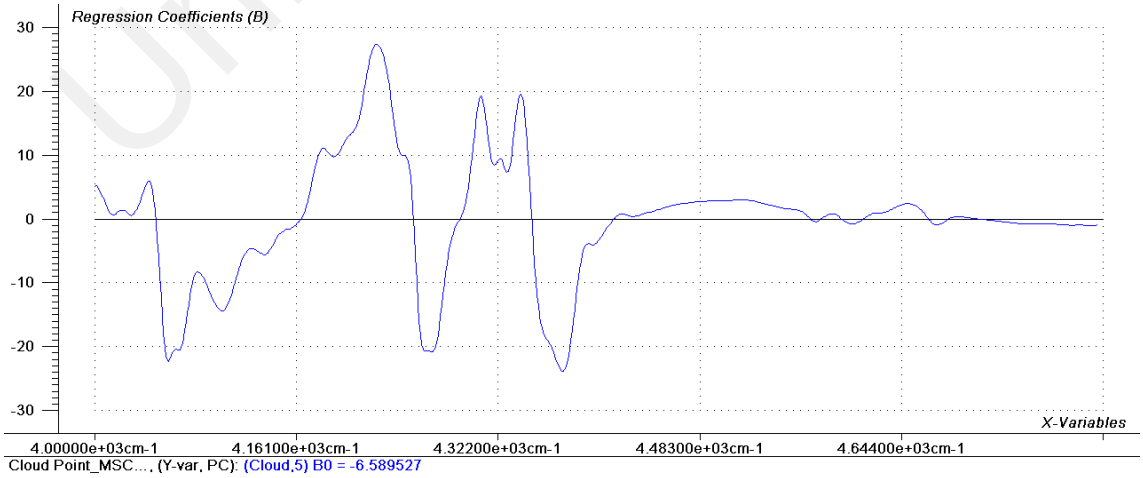


Figure 4.62: Regression coefficient plot at total 5 PCs for MSC-PCR, cloud point

(c) SGSD-PLSR (Total of 4 PCs)

The RMSE plot for the SGSD-PLSR cloud point model indicates two possible PCs, which are total of 4 PCs and total of 8 PCs. The software diagnostic tool identified total of 8 PCs as the optimum PC rather than total of 4 PCs (first minimum curve), where the RMSE value is much lower (Figure 4.63). At total of 4 PCs, the difference of RMSE for calibration and validation indicates no a significant difference compared with total of 8 PCs, which is significant difference.

Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 4 PCs. The *Y* variance explained about 89% (Figure 4.64 and 4.65).

Total of 8 PCs, although the RMSE value is lower than total of 4 PCs and *Y* variance explained (96%) is higher than total of 4 PCs (89%), much noise is indicated at both x-loading and regression coefficient plots (Figure 4.66 and Figure 4.67). It might lead to instability and inaccuracy of the model because the noise was embedded in the regression for future predictions.

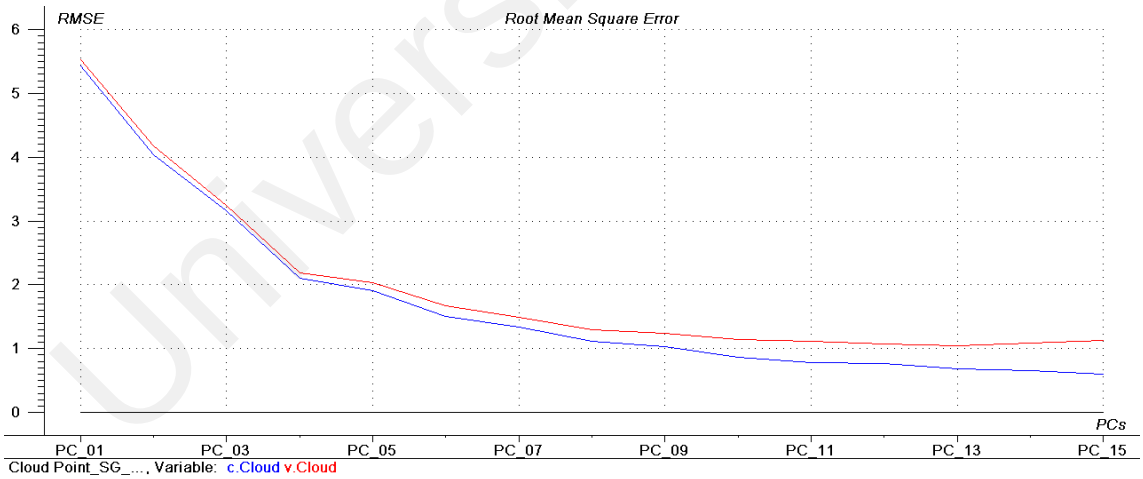


Figure 4.63: RMSE versus PCs plot for SGSD-PLSR, cloud point

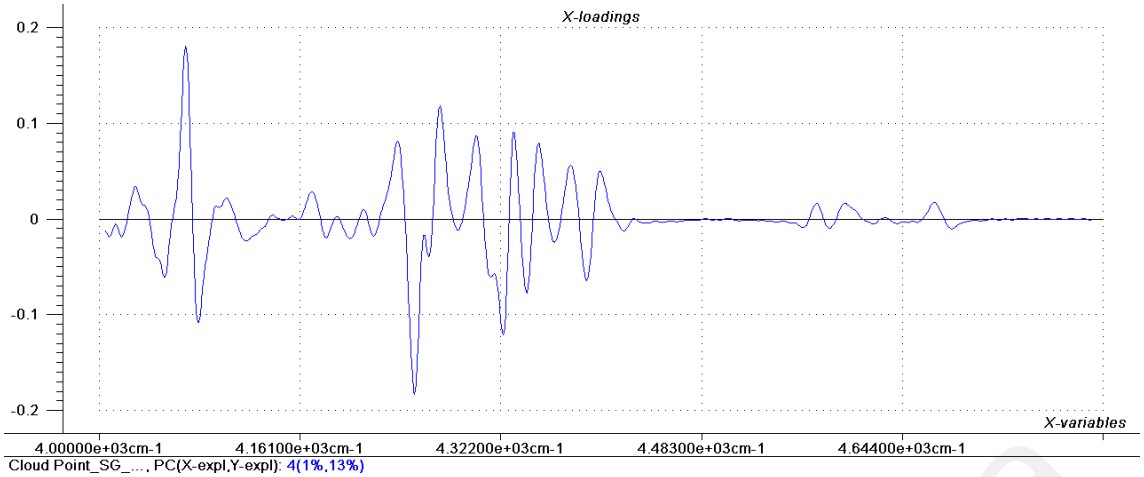


Figure 4.64: *X*-loading plot at total 4 PCs for SGSD-PLSR, cloud point

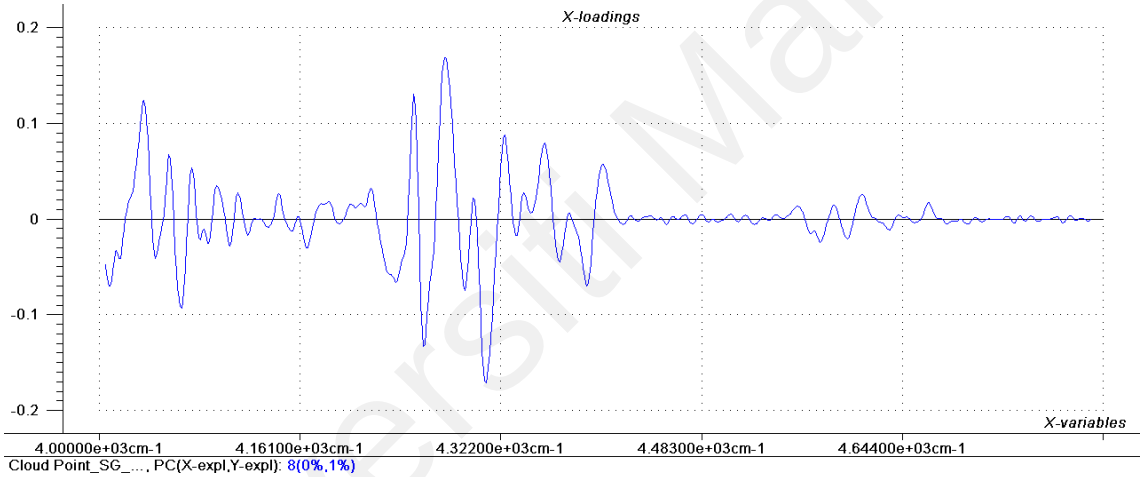


Figure 4.65: *X*-loading plot at total 8 PCs for SGSD-PLSR, cloud point

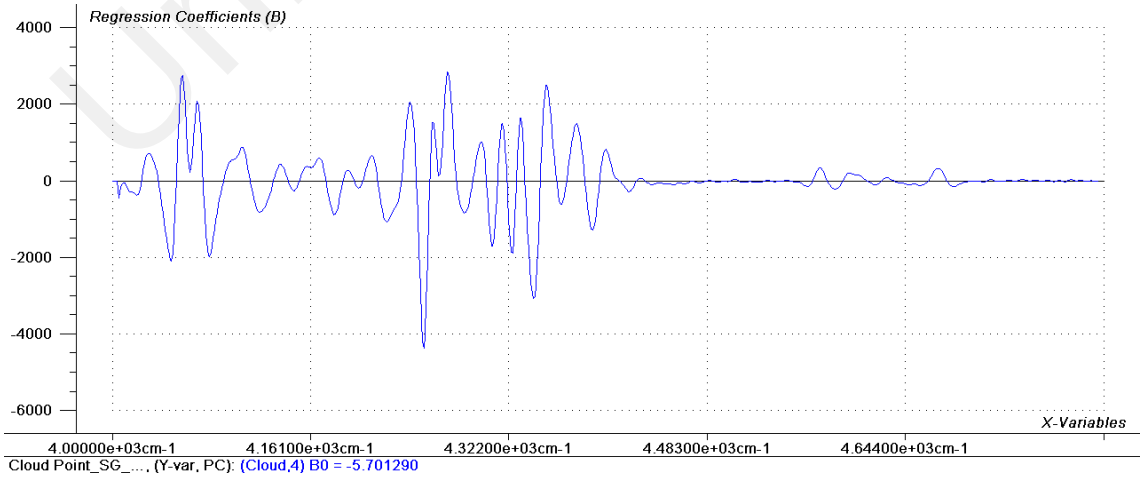


Figure 4.66: Regression coefficient plot at total 4 PCs for SGSD-PLSR, cloud point

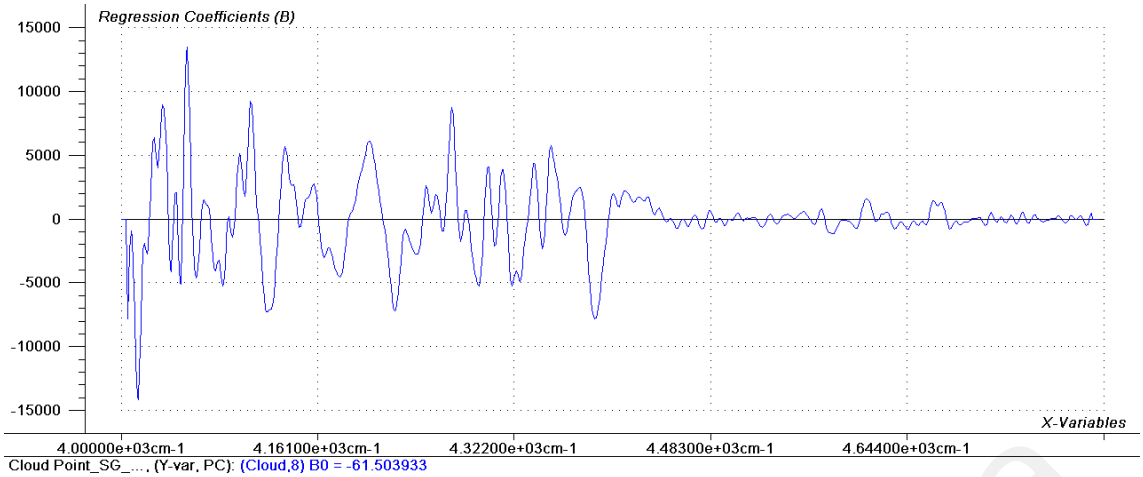


Figure 4.67: Regression coefficient plot at total of 8 PCs for SGSD-PLSR, cloud point

(d) SGSD-PCR (Total of 5 PCs)

The RMSE plot indicates a minimum curve with total of 5 PCs as the optimum PC for the MSC-PCR flash point model. The software diagnostic tool also suggested total of 5 PCs is the local minimum as well.

At total of 5 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 5 PCs. The *Y* variance explained about 92% (Figure 4.68-4.70).

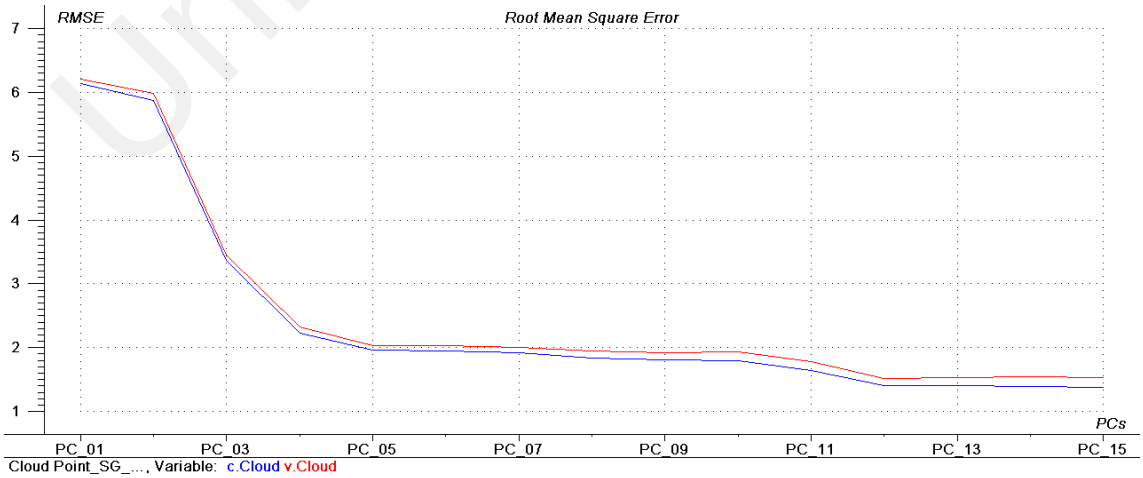


Figure 4.68: RMSE versus PCs plot for SGSD-PCR, cloud point

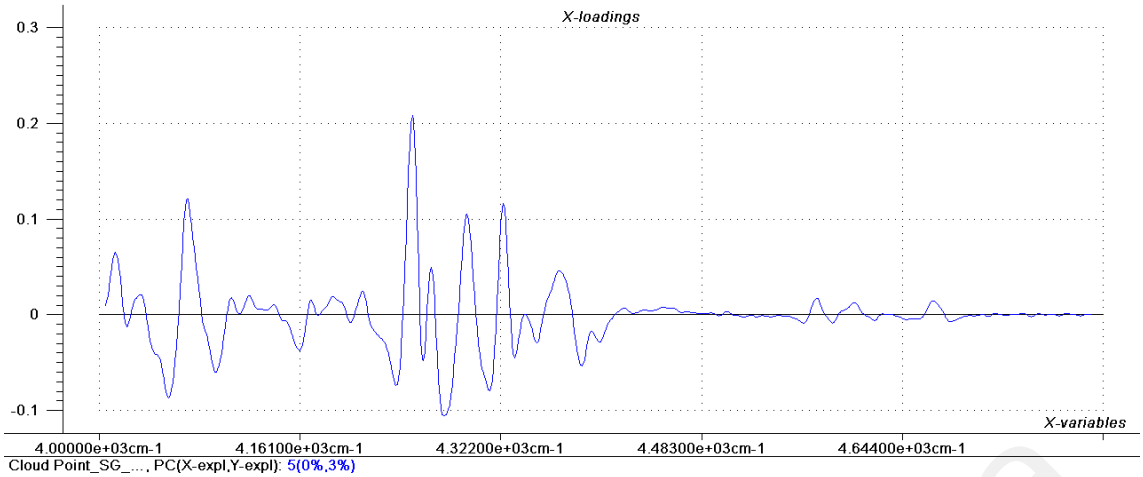


Figure 4.69: *X*-loading plot at total 5 PCs for SGSD-PCR, cloud point

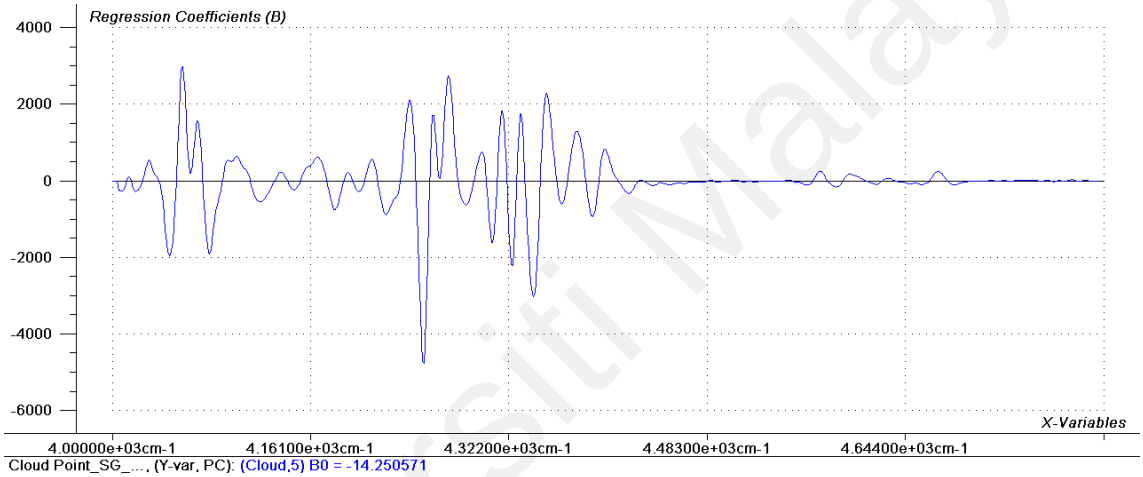


Figure 4.70: Regression coefficient plot at total 5 PCs for SGSD-PCR, cloud point

4.6.1.4 Calibration of Cetane Index

(a) MSC-PLSR (Total of 4 PCs)

The RMSE plot indicates a sharped minimum curve with total of 4 PCs as the optimum PC for the MSC-PLSR cetane index model. The software diagnostic tool also suggested total of 4 PCs is the local minimum as well.

At total of 4 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 4 PCs. The *Y* variance explained about 99% (Figure 4.71-4.73).

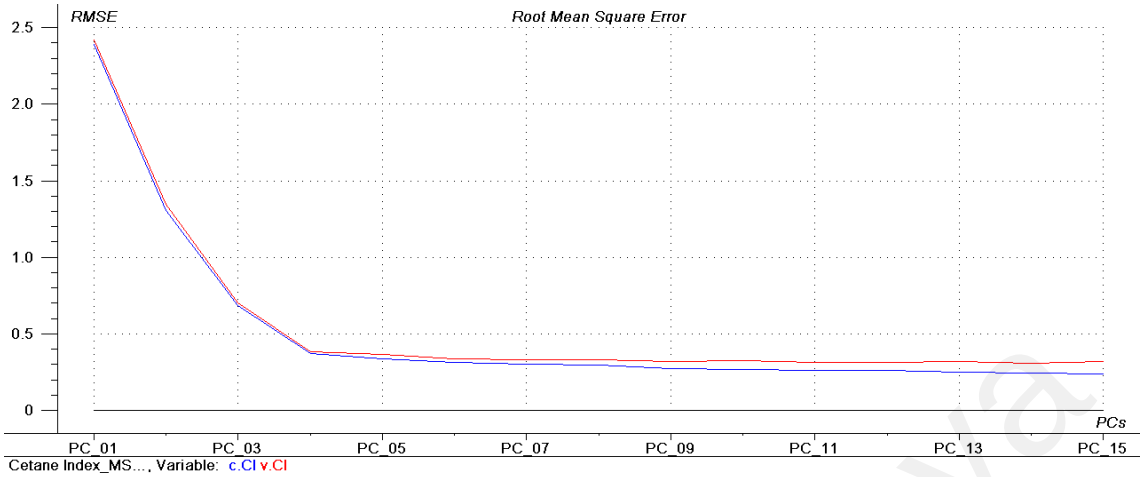


Figure 4.71: RMSE versus PCs plot for MSC-PLSR, cetane index

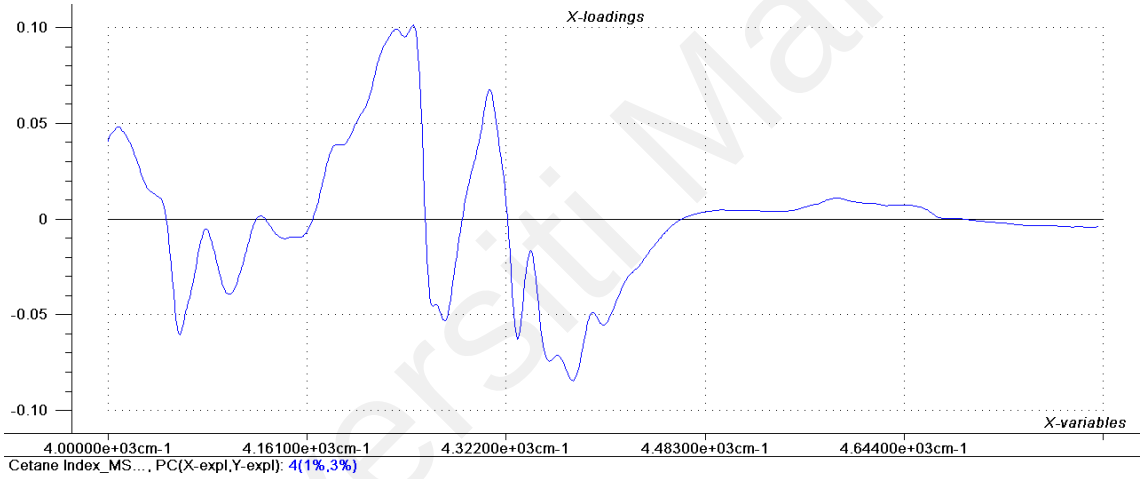


Figure 4.72: *X*-loading plot at total 4 PCs for MSC-PLSR, cetane index

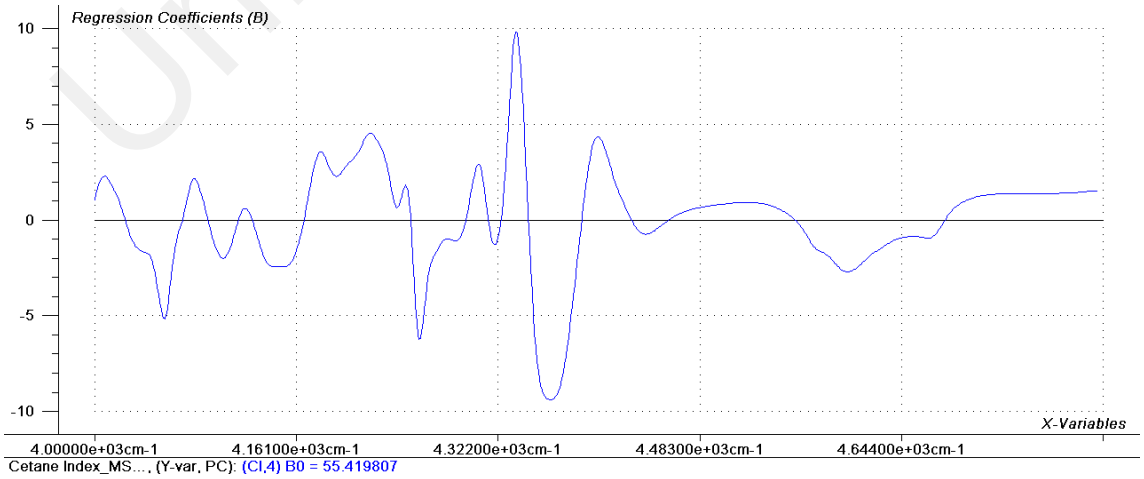


Figure 4.73: Regression coefficient plot at total 4 PCs for MSC-PLSR, cetane index

(b) MSC-PCR (Total of 4 PCs)

The RMSE plot indicates a sharped minimum curve with total of 4 PCs as the optimum PC for the MSC-PCR cetane index model. The software diagnostic tool also suggested total of 4 PCs is the local minimum as well.

At total of 4 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 4 PCs. The *Y* variance explained about 98% (Figure 4.74-4.76).

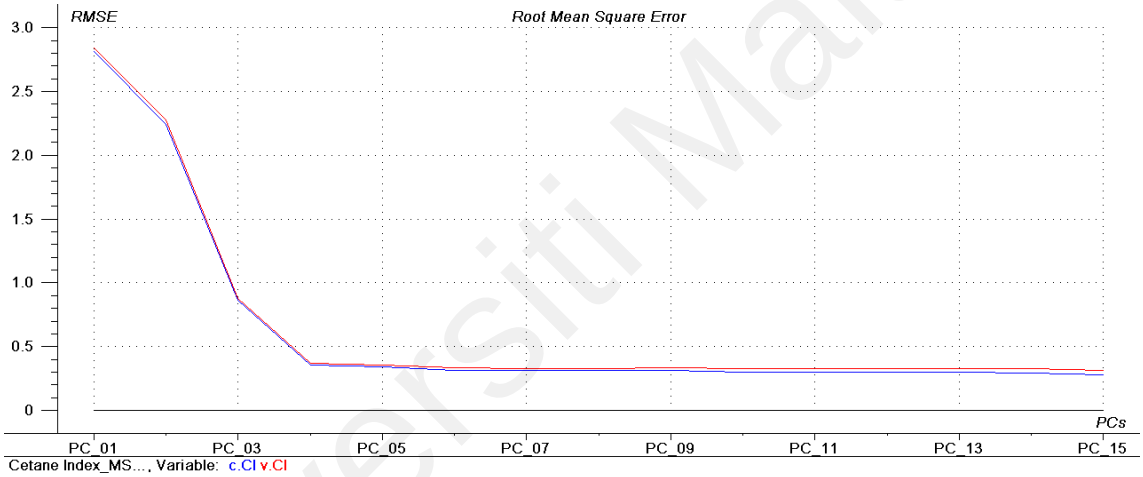


Figure 4.74: RMSE versus PCs plot for MSC-PCR, cetane index

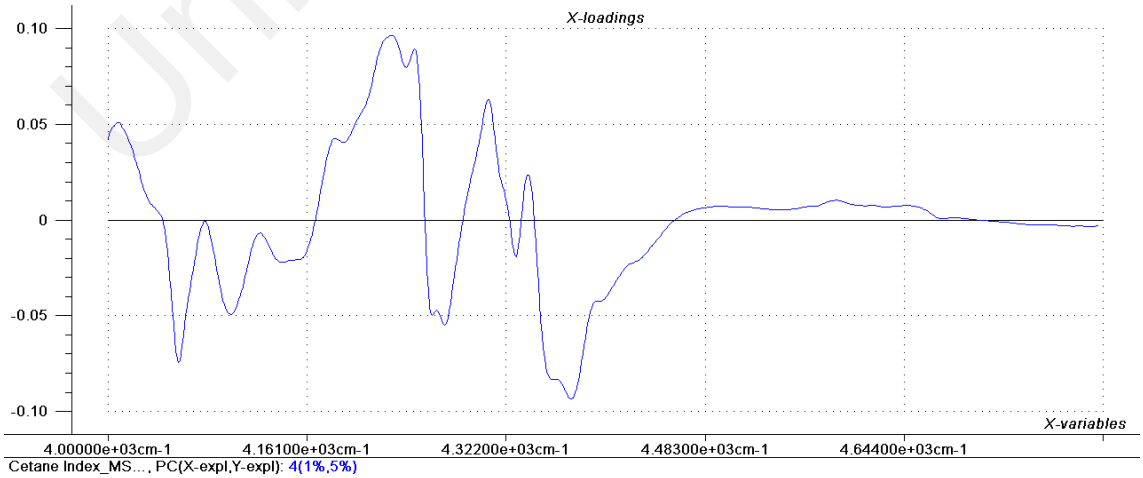


Figure 4.75: *X*-loading plot at total 4 PCs for MSC-PCR, cetane index

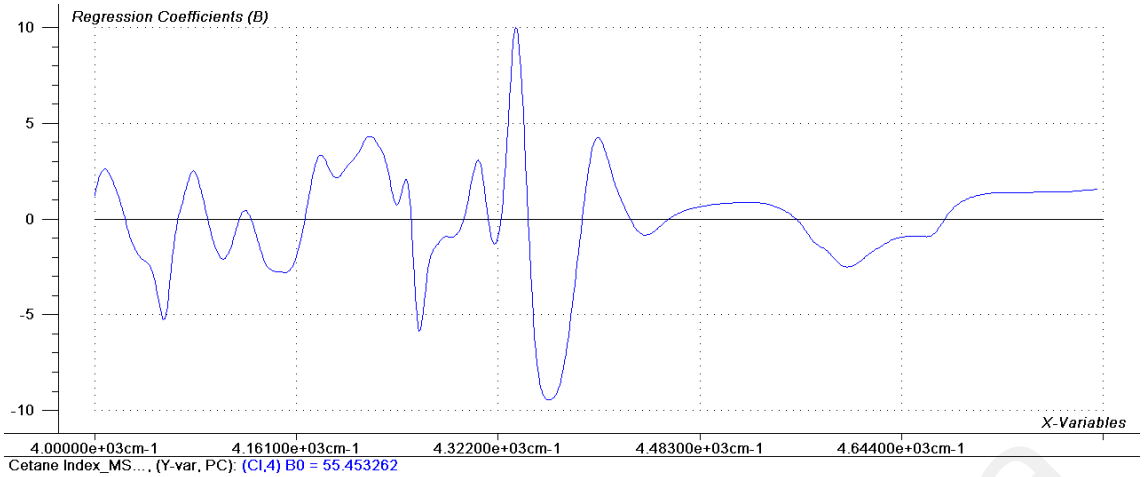


Figure 4.76: Regression coefficient plot at total 4 PCs for MSC-PCR, cetane index

(c) SGSD-PLSR (Total of 4 PCs)

The RMSE plot indicates a sharped minimum curve with total of 4 PCs as the optimum PC for the SGSD-PLSR cetane index model. The software diagnostic tool also suggested total of 4 PCs is the local minimum as well.

At total of 4 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 4 PCs. The *Y* variance explained about 99% (Figure 4.77-4.79).

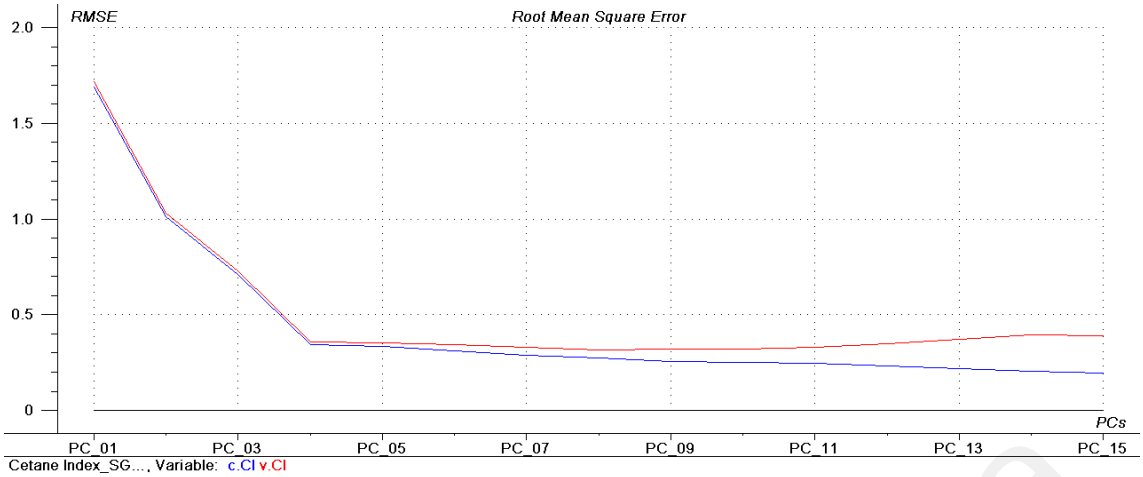


Figure 4.77: RMSE versus PCs plot for SGSD-PLSR, cetane index

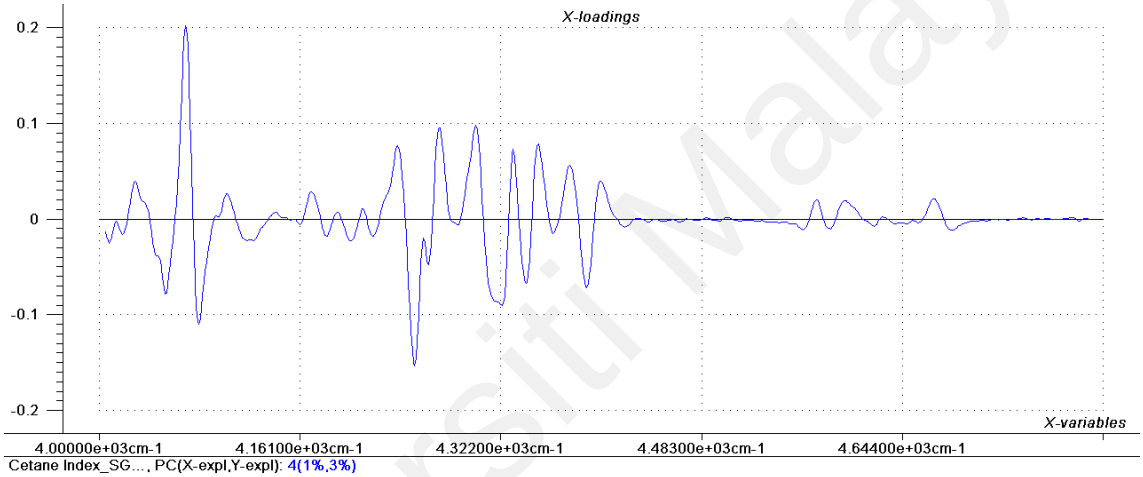


Figure 4.78: *X*-loading plot at total 4 PCs for SGSD-PLSR, cetane index

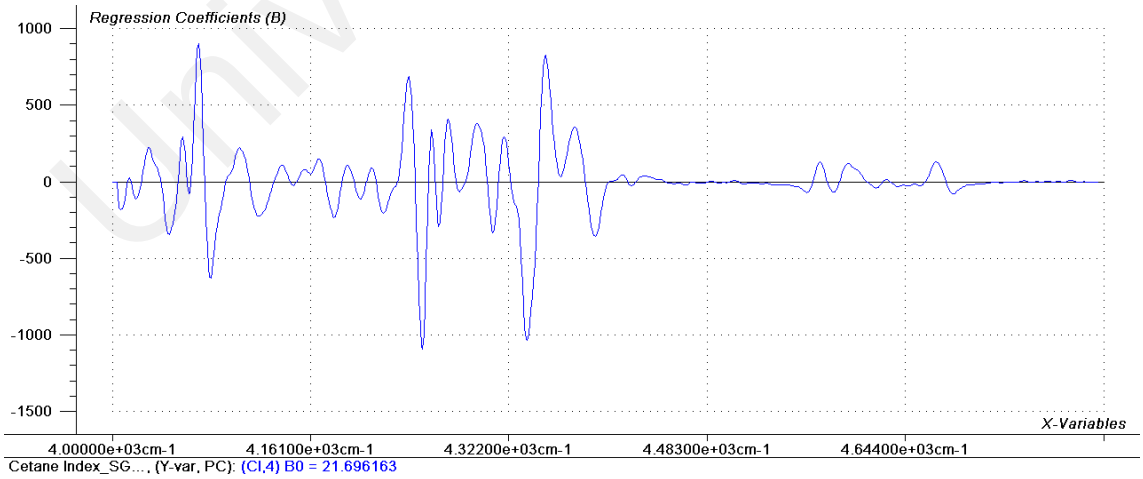


Figure 4.79: Regression coefficient plot at total 4 PCs for SGSD-PLSR, cetane index

(d) SGSD-PCR (Total of 4 PCs)

The RMSE plot indicates a sharped minimum curve with total of 4 PCs as the optimum PC for the SGSD-PLSR cetane index model. The software diagnostic tool also suggested total of 4 PCs is the local minimum as well.

At total of 4 PCs, the difference of RMSE for calibration and validation indicates an insignificant difference. Further evaluation of the x-loading and regression coefficient plots indicated no noise at total of 4 PCs. The *Y* variance explained about 99% (Figure 4.80-4.82).

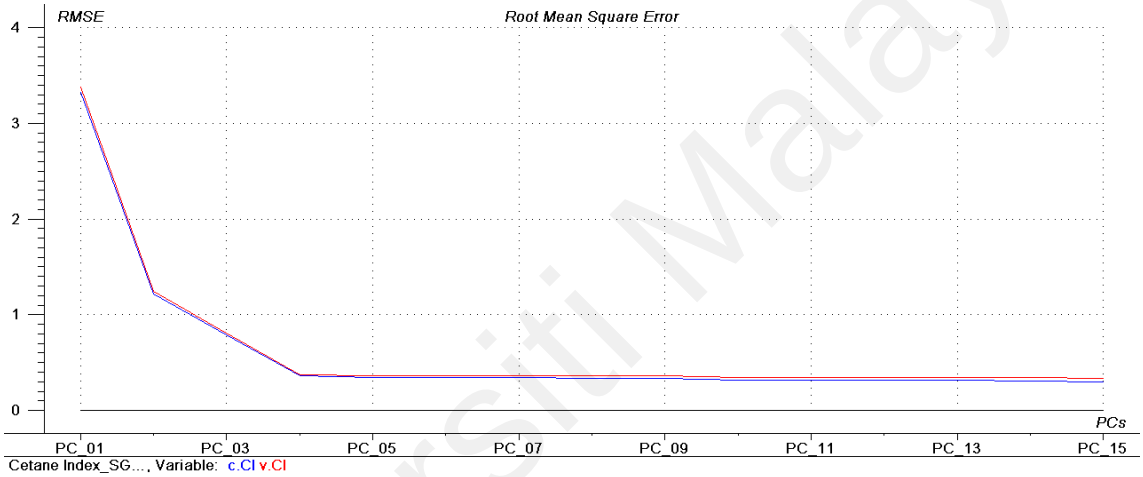


Figure 4.80: RMSE versus PCs plot for SGSD-PCR, cetane index

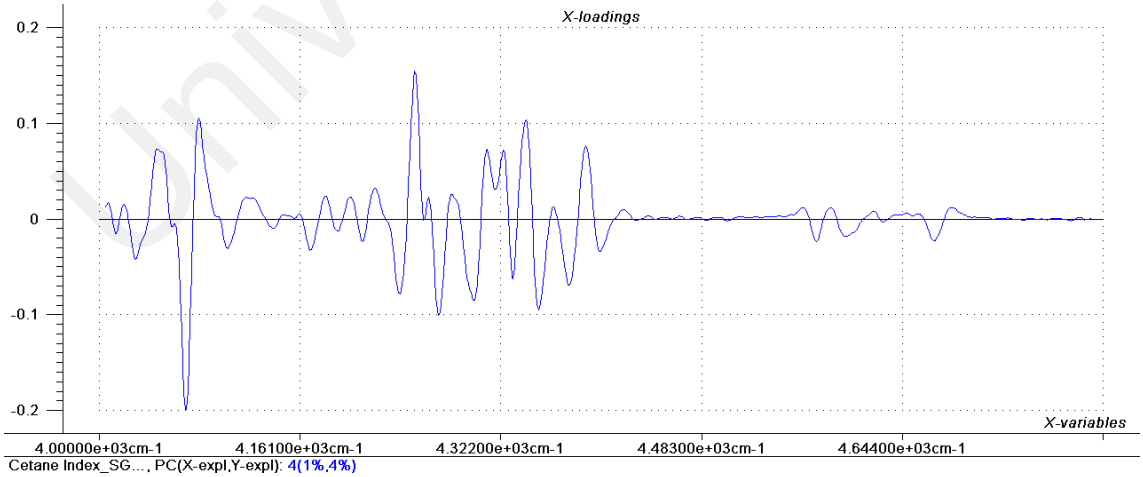


Figure 4.81: *X*-loading plot at total 4 PCs for SGSD-PCR, cetane index

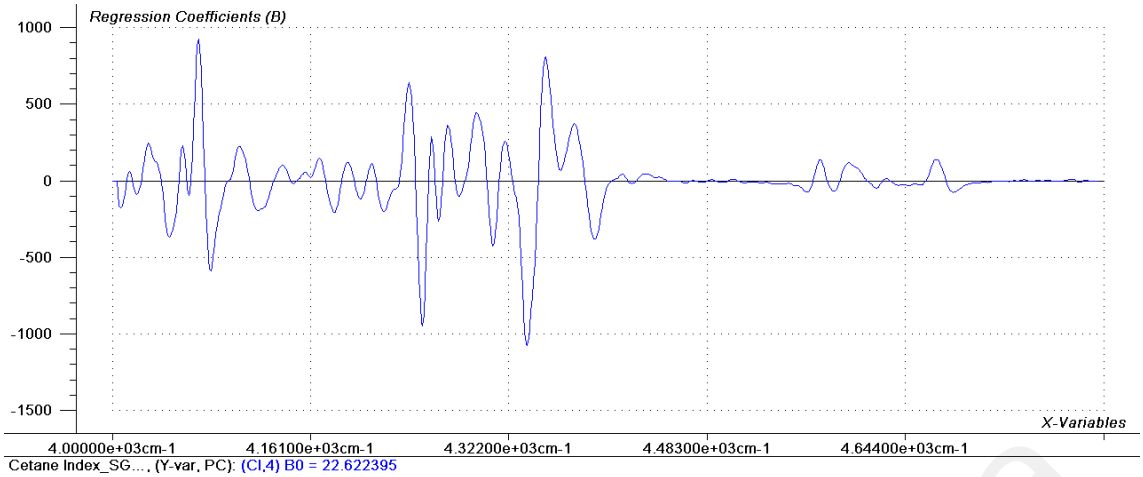


Figure 4.82: Regression coefficient plot at total 4 PCs for SGSD-PCR, cetane index

4.6.2 Diesel modelling

4.6.2.1 Modelling the boiling point at 95% recovery

Diesel fuel comprises hydrocarbon derivatives, including paraffin, naphthene, olefins, and aromatics. In general, paraffin has a lower boiling point than naphthenic and aromatic compounds of the same carbon number; together, they express the boiling range distribution (Coker, 2018). Hence, the NIR combination bands region (4800-4000 cm^{-1}) that donated by the combinational of vibrational modes of the C-H bond of methyl, methylene, and aromatic rings are responsive to the compositional variations (i.e. the proportion of medium and heavy oil fractions in the diesel samples), and would facilitate the modelling of boiling point at 95% recovery temperature.

The predictive performance of NIR-T95% models derived is listed in Table 4.2. Considering the calibration range of 305 ºC to 372 ºC, all the obtained models exhibit excellent RMSECV, RMSEP, R^2, RPD and satisfy the reproducibility requirement (7.5 ºC) of the reference method (ASTM D86-20, 2020). The comparable performance among the linear models (PCR and PLSR) reflected the fitness of the selected spectral range and partitioned calibration/validation

sets in the NIR-T95% modelling in addition to the spectral pre-processing methods. The slight difference observed between each RMSECV and RMSEP pair set aside the concern of false-positive prediction (Hradecká et al., 2021); and the respective bias was also found statistically insignificant at a significance level of 0.05. In terms of precision, the RPD values were always greater than 3. According to Rossel et al. (2006), RPD values can be generally classified into six classes: excellent (>2.5), very good (2.5-2.0), good (2.0-1.8), fair (1.8-1.4), poor (1.4-1.0), and very poor (<1.0). Based on the validation data, MSC-PLSR model was found to be the best for precise measurement of T95 of routine diesel samples.

From Table 4.2, it is clearly demonstrated that the prediction capability of the developed models is comparable with those reported advances using two-stage sequential spectral pre-processing strategy (Gonzaga & Pasquini, 2010; Palou et al., 2017). This acknowledged the dependency of model performance based upon the overall calibration strategy against targeted samples instead of solely the signal-processing and/or modelling algorithm.

Table 4.2: Comparison of the model performances in NIR determination of T95

Pre-processing	Model	R^2	LVs	RMSECV	RMSEP	RPD	References
MSC	PLSR	0.969	7	4.04	3.49	5.38	This study
MSC	PCR	0.935	4	5.43	5.19	3.73	
SG-SD	PLSR	0.945	4	5.26	4.72	4.02	
SG-SD	PCR	0.942	5	5.33	4.92	3.81	
SG-SD + SNV	PLSR	-	8	-	5.39	-	(Palou et al., 2017)
SD-GS + SNV	PLSR	-	8	-	5.30	-	
SG + MC	PLSR	0.84	5	4.7	5.0	-	(Gonzaga & Pasquini, 2010)

-Not reported; SNV (standard normal variate); SD-GS (second derivative-gap segment); MC (mean centering)

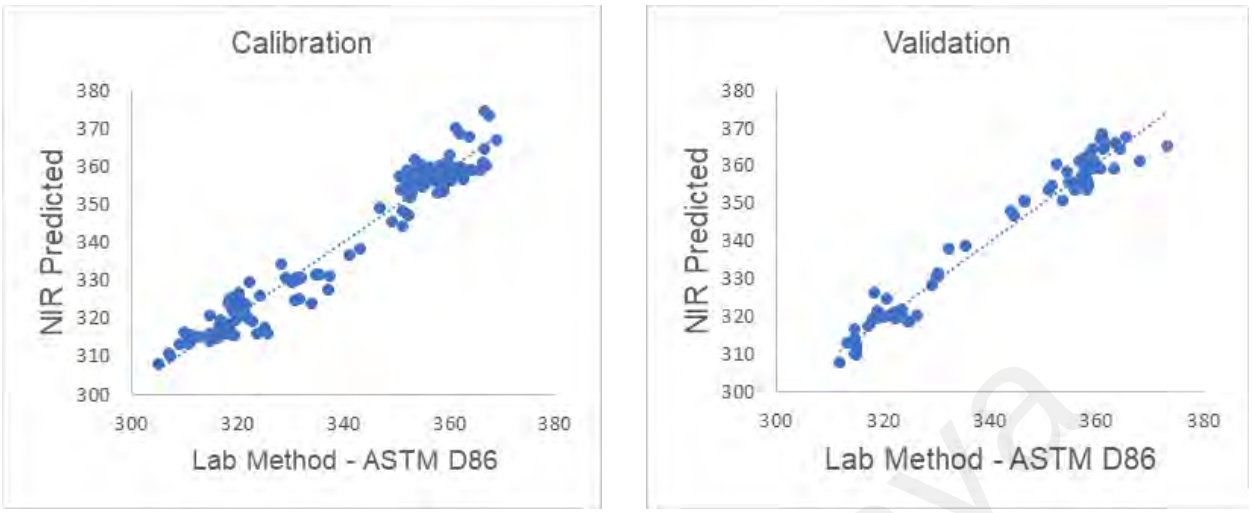


Figure 4.83: T95 Distillation MSC-PLSR Model at total of 7 LVs

4.6.2.2 Modelling the Flash Point

In-process control and optimisation, the flash point (FP) of diesel product is adjusted to meet the specification requirement by regulating the blending component ratio between the light and the heavy hydrocarbon fractions. This suggested that the combination bands region (4800-4000 cm^{-1}) adopted in NIR-T95% modelling would also be appropriate for simultaneously determining FP (Alves et al., 2012).

Table 4.3 summarises the performance characteristics established for the NIR models under current experimental settings compared to other reported models, including a model by support vector machine regression (SVMR). From the data, the developed models (that calibrated for FP between 62 and 92 ^{o}C) showed improved predictive performance in terms of RMSEP particularly; and satisfied the ASTM D93-20 (2020) specifications on test method producibility (5.44 ^{o}C) and repeatability (2.22 ^{o}C). At a significance level of 0.05, the measurement bias was negligible. When considering the FP at the Malaysian regulatory limit, i.e. a minimum value of 60.0 ^{o}C, a relative

prediction error of around 3.65% could be achieved with the SG-SD-PLSR model that had explained 97% of the FP variance with 5 LVs. These NIR models would be useful for the online control of crude distillation unit process via real-time FP monitoring so as to optimise the cut point between front-end diesel and back-end kerosene fractions.

Table 4.3: Comparison of the model performances in NIR determination of FP

Pre-processing	Model	R^2	LVs	RMSECV	RMSEP	RPD	References
MSC	PLSR	0.950	9	1.71	1.50	4.29	This study
MSC	PCR	0.892	5	2.25	2.02	3.01	
SG-SD	PLSR	0.949	5	1.74	1.39	4.25	
SG-SD	PCR	0.912	4	2.03	1.78	3.18	
SG-SD	PLSR	-	7	-	3.72	-	(Palou et al., 2017)
SD-DS	PLSR	-	3	-	3.68	-	
SNV	PLSR	0.698	-	-	3.77	-	(Alves, Henriques, & Poppi, 2012)
SNV	SVMR	0.936	-	-	1.98	-	

-Not reported; SNV (standard normal variate); SVMR(support vector machine regression)

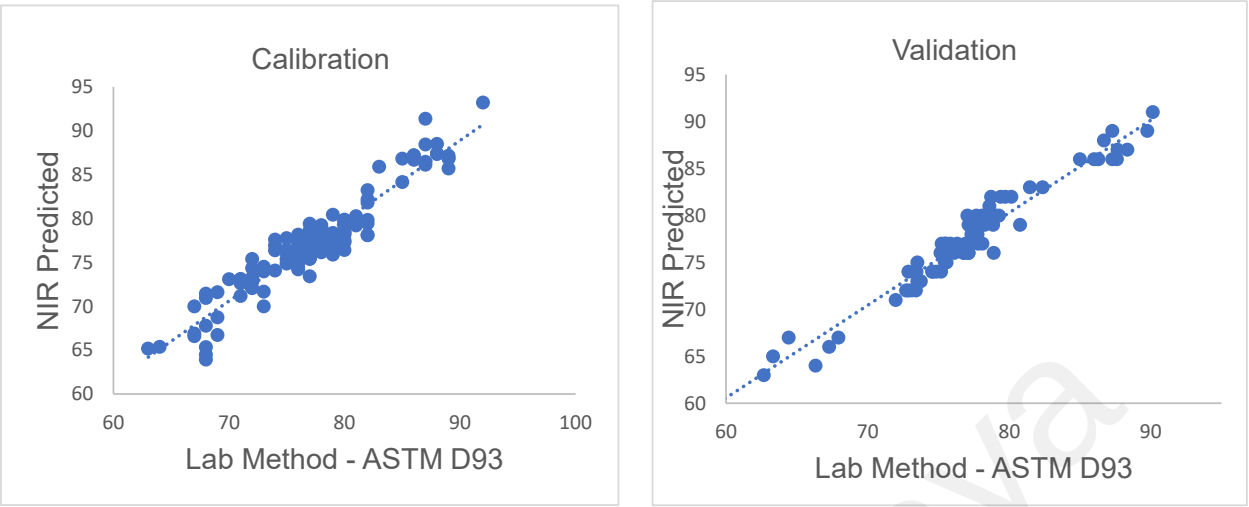


Figure 4.84: Flash Point SG-SD-PLSR Model at total of 5 LVs

4.6.2.3 Modelling the Cloud Point

Cloud Point (CP) refers to the minimum temperature at which the diesel fuel undergoes phase separation. The tendency of becoming cloudy at low temperatures is controlled by the agglomeration of paraffinic hydrocarbon at the molecular level (Adebiyi, 2020). Thus, the aforementioned NIR absorption range adopted for spectrometric measurement of temperature properties of diesel could also be used for CP determination.

The performance characteristics of the developed NIR-CP models are given in Table 4.4. Referring to the calibration and validation results between -21 ^{o}C and 14 ^{o}C, the developed models showed satisfactory performances in terms of prediction bias ($p>0.05$), precision ($RPD>2.5$), and met the ASTM D2500-17 (2017) requirement of reproducibility. Likewise, the linear models either PCR or PLS are sufficient for NIR-CP calibration and prediction where the noise and potential signal suppression had been addressed by the spectral region selection and pre-processing.

Table 4.4: Comparison of the model performances in the determination of CP

Pre-processing	Model	R^2	LVs	RMSECV	RMSEP	RPD	References
MSC	PLSR	0.955	7	1.68	1.53	4.66	This study
MSC	PCR	0.893	5	2.34	2.36	2.93	
SG-SD	PLSR	0.915	4	2.19	2.14	3.28	
SG-SD	PCR	0.919	5	2.02	2.12	3.29	
SG-SD	PLSR	-	5	-	1.22	-	(Palou et al., 2017)
SD-DS	PLSR	-	5	-	1.39	-	
-	PLSR	-	-	-	1.80	-	(Process Insight, 2019)

-Not reported; SD-GS (second derivative-gap segment)

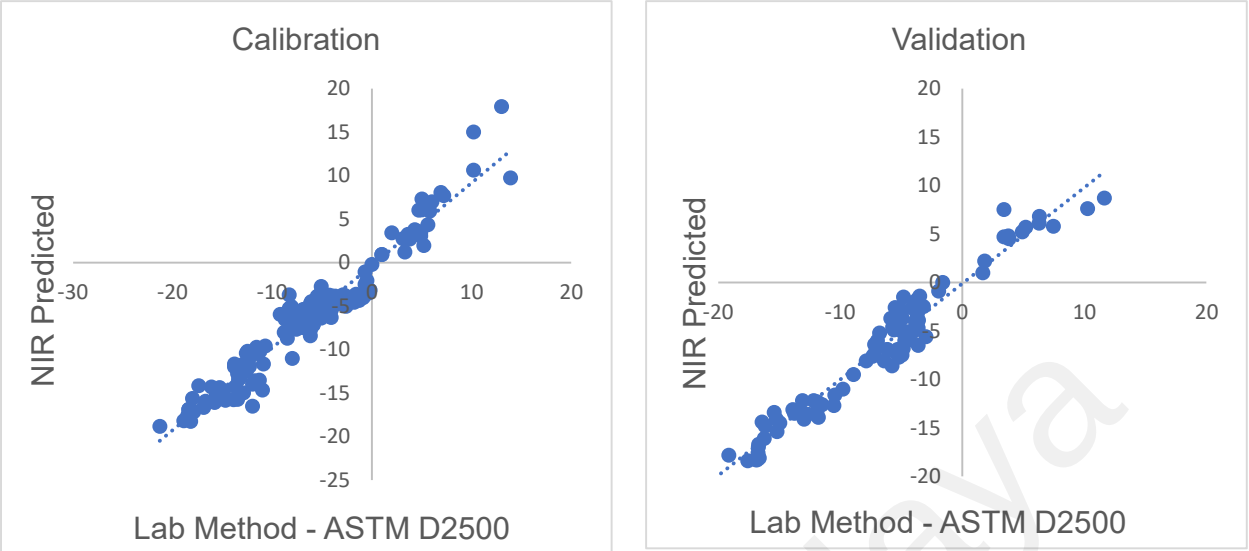


Figure 4.85: Cloud Point MSC-PLSR Model at total of 7 LVs

4.6.2.4 Modelling of Cetane Index

Cetane number is a diesel combustion indicator related to the compositional variation in paraffinic and aromatic hydrocarbons (ASTM D613-03, 2003). Alternatively, it can be computed in terms of Cetane Index (CI) in the absence of a test engine by using the diesel density and the boiling temperature (ASTM D976-06, 2006). Based on such arguments, the CI is, therefore, ready to be estimated via a multivariate NIR calibration strategy.

Table 4.5 lists the multivariate NIR models in CI estimation; for this study, the calibration range was from 45-63. Apparently, CI can be spectrometric determined using various calibration strategies, where an appropriate combination of data analytics could deliver promising results. For example, advanced machine learning tools, i.e. artificial neural networks (ANN) and support vector machine regression (SVMR), have been reported together with diverse signal pre-processing algorithms. As shown in Table 4.5, the models obtained from this study fit online CI

measurement, with an excellent degree of precision and unbias compared to other reported models, including the ASTM reference method (ASTM D976-06 2006). Notably, the SG-SD-PLSR model that explained 98% of the CI variation with 4 LVs showed the best predictive performance. When considering CI measurement at the limit specified in the local requirement of diesel fuel, i.e. 47, the relative prediction error of the model was 0.77%.

Table 4.5: Comparison of the model performances in the determination of CI

Pre-processing	Model	R^2	LVs	RMSECV	RMSEP	RPD	References
MSC	PLSR	0.9877	4	0.39	0.30	12.02	This study
MSC	PCR	0.9887	4	0.37	0.30	11.90	
SG-SD	PLSR	0.9898	4	0.36	0.27	13.35	
SG-SD	PCR	0.9887	4	0.38	0.31	11.40	
ISPXY+ IGWO	SVMR	0.936	-	-	1.57	-	(Liu et al., 2022)
SG-FD + CC	SVMR	0.887	-	-	1.43	-	(Wang et al., 2020)
SG-FD + SCARS	SVMR	0.936	-	-	0.20	-	
SG-SD	PLSR	-	2	-	1.02	-	(Palou et al., 2017)
SD-DS	PLSR	-	4	-	1.02	-	
MSC + LS-SVM + DOSC	PLSR	-	-	0.56	0.24	-	(Feng, Wu, & Zeng, 2015)
SNV	PLSR	0.894	-	-	0.56	-	(Alves et al., 2012)
SG + MC	PLSR	0.94	2	0.4	0.5	-	(Gonzaga & Pasquini, 2010)
FD + VN	ANN	0.971	-	0.20	0.44	-	(Santos Jr et al., 2005)

-Not reported; ANN (artificial neural network); ISPXY (improved XY co-occurrence distance); IGWO (improved grey wolf optimization); SVMR(support vector machine regression); SG-FD (Savitzky-Golay first derivative); CC(correlation coefficient); SCARS (stability competitive adaptive reweighted sampling); SD-GS (second derivative-gap segment); LS-SVM (least square support vector machine); DOSC (direct orthogonal signal correction); FD (first derivative); VN (vector normalization).

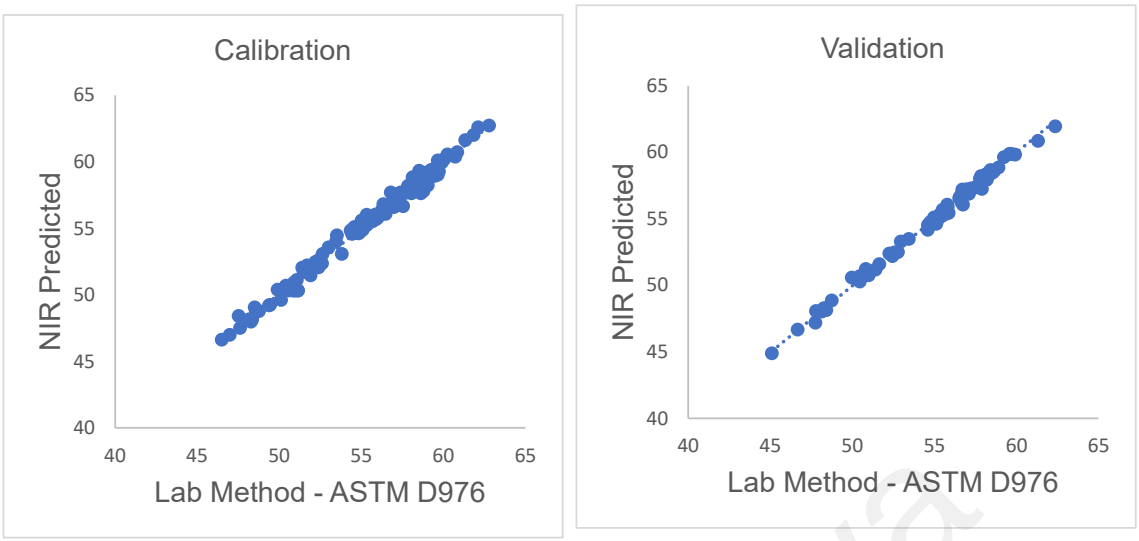


Figure 4.86: Cetane Index SG-SD-PLSR Model at total of 4 LVs

CHAPTER 5: CONCLUSION

Three types of petroleum products (gasoline, diesel and kerosene) were well discriminated against using FT-NIR spectroscopy. With the combination of PCA, the unknown petroleum products will be identified based on the similarities of their spectral features and compositions. As a result, FT-NIR coupled with PCA, fast qualitative measurement of less than a minute, and cheaper cost can be applied without the laboratory reference method analysis.

In the case study for differentiation between gasoline with and without additives, the PCA analysis outcomes provided not-so-clear discrimination and classification between them due to the concentration of additives being at a low concentration (in ppm(wt)) level compared with the compositions of the gasoline, which at the % volume level. However, the FT-NIR spectrum in the region of 4100-4000 cm^{-1} and 4700-4500 cm^{-1} indicates slight differentiation between gasoline and without additives.

In diesel, the differentiation and classifications were clearly separated into two clusters, i.e., diesel with and without PME, with a high degree of accuracy from the PCA outcomes. The result is expected because the composition varies depending on the presence of methyl ester groups (C=O and C-O) and the concentrations of PME blended into diesel with the range of 7 vol/vol % up to 21 vol/vol% compared with diesel without PME blended.

Determination of fuel properties is vital for process control and optimisation of diesel blending, in addition to complying with quality specifications and emission standards. By appropriate spectral region selection, calibration/validation set partition, spectra pre-processing, and regression algorithm, this study has delivered an FT-NIR alternative for simultaneous determination of T95, FP, CP and CI, where the prediction performance is comparable to the ASTM reference methods. Since the targeted properties strongly depend

upon hydrocarbon variability in the diesel sample, the NIR combination region (4000- 4800 cm^{-1}) granted vital information at the molecular level for mathematical modelling. At the same time, the calibration/validation set partition via random selection provided a strong representation throughout the working region, which was evaluated by the score plots in a reduced dimension. These two steps donated the fundamentals for NIR calibration; thus, comparable models were achieved by MSC/SG-SD and PCR/PLSR algorithms. This revealed that the model performance relies upon every component along the calibration process; appropriate pre-treatment is crucial to deliver to a model that fits its purpose.

REFERENCES

Abbisek Ukil, IEEE Jakb Bernasconi, Hubert Braendle, Henry Buijs, Sacha Bonenfant (2010). Improved calibration of near infrared spectra by using ensembles of neural network models. *IEEE Sensors Journal,* volume 10(3): 578-584.

Adebiyi, F. M. (2020). Paraffin wax precipitation/deposition and mitigating measures in oil and gas industry: a review. *Petroleum Science and Technology*, 38(21), 962-971.

Adesogan, A., Owen, E., & Givens, D. (1998). Prediction of the in vivo digestibility of whole crop wheat from in vitro digestibility, chemical composition, in situ rumen degradability, in vitro gas production and near infrared reflectance spectroscopy. *Animal feed science and technology,* 74(3), 259-272.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis,* 22(5), 717-727.

Aitani, A. M. (2004). Oil refining and products. *Encyclopedia of energy,* 4, 715-729.

Aleme, H.G. and Barbeira, P.J.S. (2012). Determination of flash point and cetane index in diesel using distillation curves and multivariate calibration. *Fuel* 102: 129-134.

Altgelt, K. H. (1993). Composition and analysis of heavy petroleum fractions: *CRC press*.

Alves, J.C.L., Henriques, C.B., and Poppi, R.J. (2012). Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimiser system. *Fuel* 97: 710-717.

André, M. P., Janée, H. S., Martin, P. J., Otto, G. P., Spivey, B. A., & Palmer, D. A. (1997). High-speed data acquisition in a diffraction tomography system employing large-scale toroidal arrays. *International Journal of Imaging Systems and Technology*, 8(1), 137-147.

ASTM D1298-12b (2017). Standard test method for density, relative density, or API gravity of crude petroleum and liquid petroleum products by hydrometer method, ASTM International: West Conshohocken, PA.

ASTM D2500-17 (2017). Standard test method for cloud point of petroleum products and liquid fuels. ASTM International: West Conshohocken, PA.

ASTM D613-03 (2003). Standard test method for cetane number of diesel fuel oil. ASTM International: West Conshohocken, PA.

ASTM D86-20 (2020). Standard test method for distillation of petroleum products and liquid fuels at atmospheric pressure. ASTM International: West Conshohocken, PA.

ASTM D93-20 (2020). Standard test methods for flash point by Pensky-Martens closed cup tester. ASTM International: West Conshohocken, PA.

ASTM D975-21 (2021). Standard specification for diesel fuel oils. ASTM International: West Conshohocken, PA.

ASTM D976-06 (2016). Standard test method for calculated cetane index of distillate fuels, ASTM International: West Conshohocken, PA.

ASTM E1655-00 (2000). Standard practices for infrared multivariate quantitative analysis. ASTM International: West Conshohocken, PA.

Bahadur, N. P., Boocock, D. G., & Konar, S. K. (1995). Liquid hydrocarbons from catalytic pyrolysis of sewage sludge lipid and canola oil: evaluation of fuel properties. *Energy & fuels,* 9(2), 248-256.

Baibing Li, Elaine Martin and Julian Morris (2001). Latent variables selection in partial least squares modelling. *IFAC Publications* : 463-468.

Balaji, G. N., Suriya, N. H., AnandVikash, S., Arun, R., & Kumar, S. A. (2017). Analysis of Various Liquid Components under Different Temperature and Density Constraints Pertaining To Fractional Distillation. *Imperial Journal of Interdisciplinary Research (IJIR)* Vol, 3, 664-669.

Barabas, I., Todoruţ, A., & Băldean, D. (2010). Performance and emission characteristics of an CI engine fueled with diesel–biodiesel–bioethanol blends. *Fuel,* 89(12), 3827-3832.

Barra, I., Kharbach, M., Qannari, E.M., Hanafi, M., Cherrah, Y., and Bouklouze, A. (2020). Predicting cetane number in diesel fuels using FTIR spectroscopy and PLS regression. *Vibrational Spectroscopy* 111: 103157.

Bediaga, H., Moreno, M.I., Arrasate, S., Vilas, J.L., Orbe, L., Unzueta, E., Mercader, J.P., and González-Díaz, H. (2022). Multi-output chemometrics model for gasoline compounding. *Fuel* 310: 122274.

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

Broeke, J., & Koster, T. (2019). Spectroscopic methods for online water quality monitoring. ICT for Smart Water Systems: Measurements and Data Science, 283-314.

Brown, C. H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics,* 143-155.

Bukkarapu, K.R., and Krishnasamy A. (2021). A relative assessment of chromatographic and spectroscopic based approaches to predict engine fuel properties of biodiesel. *Fuel Processing Technology* 222: 106960.

Bunaciu, A. A., Aboul-Enein, H. Y., & Hoang, V. D. (2015). Vibrational spectroscopy used in polymorphic analysis. *TrAC Trends in Analytical Chemistry*, 69, 14-22.

Burgard, D. R., & Kuznicki, J. T. (2018). Chemometrics: chemical and sensory data. CRC Press.

Chang, C., Wei, M., & Ji, H. (2020). Experimental Research on Cold Start of PFI Two-Stroke Spark-Ignition Kerosene Engine. *Journal of Energy Engineering,* 146(4), 04020030.

Chong, C., Ni, W., Ma, L., Liu, P., & Li, Z. (2015). The use of energy in Malaysia: Tracing energy flows from primary source to end use. *Energies,* 8(4), 2828-2866.

Chung, H. (2007). Applications of near-infrared spectroscopy in refineries and important issues to address. *Applied Spectroscopy Reviews,* 42(3), 251-285.

Chung, H., Ku, M.-S., & Lee, J.-S. (1999). Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene. *Vibrational spectroscopy,* 20(2), 155-163.

Claudete Fernandes Pereiraa, Maria Fernanda Pimentel, Roberto Kawakami Harrop Galvao, Fernanda Araujo Honorato, Luiz Stragevitcha, Marcelo Nascimento Martins (2008). A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers. *Analytica Chimica Acta,* volume 611: 41-47.

Coker, A. K. (2018). Petroleum Refining Design and Applications Handbook, Volume 1. John Wiley & Sons.

Currie, L. A. (1995). Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). *Pure and applied chemistry,* 67(10), 1699-1723.

da Silva, A. R., Pastore, T. C. M., Braga, J. W. B., Davrieux, F., Okino, E. Y. A., Coradin, V. T. R., Do Prado, A. G. S. (2013). Assessment of total phenols and extractives of mahogany wood by near infrared spectroscopy (NIRS). *Holzforschung,* 67(1), 1-8.

Das, K., Bajia, S. C., Sharma, J. C., Singh, S. P., Malhan, M., Kumar, S., & Das, D. K. (2022). Improvement in cold flow properties of diesel fuel by changing its composition: a case study. *Petroleum Science and Technology*, 1-12.

De Lira, L. d. F. B., De Vasconcelos, F. V. C., Pereira, C. F., Paim, A. P. S., Stragevitch, L., & Pimentel, M. F. (2010). Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. *Fuel*, 89(2), 405-409.

De Maesschalck, R., Candolfi, A., Massart, D., & Heuerding, S. (1999). Decision criteria for soft independent modelling of class analogy applied to near infrared data. *Chemometrics and intelligent laboratory systems,* 47(1), 65-77.

de Paulo, E.H., dos Santos, F.D., Folli, G.S., Santos, L.P., Nascimento, M.H.C., Moro, M.K., da Cunha, P.H.P., Castro, E.V.R., Neto, A.C., and Filgueiras, P.R. (2021). Determination of gross calorific value in crude oil by variable selection methods applied to 13C NMR spectroscopy. *Fuel:* 122527.

de Souza, D.C.M., Cabrita, L., Galinha, C.F., and Reis, M.S. (2021a). PAT soft sensors for wide range prediction of key properties of diesel fuels and blending

components for the oil industry. *Computers & Chemical Engineering* 153: 107449.

de Souza, D.C.M., Cabrita, L., Galinha, C.F., Rato, T.J., and Reis, M.S. (2021b). A Spectral AutoML approach for industrial soft sensor development: Validation in an oil refinery plant. *Computers & Chemical Engineering* 150: 107324.

Dehaghania A.H.S. and Rahimi, R. (2019). An experimental study of diesel fuel cloud and pour point reduction using different additives. *Petroleum* 5(4): 413-416.

Demirbas, A., Alidrisi, H., & Balubaid, M. (2015). API gravity, sulfur content, and desulfurization of crude oil. *Petroleum Science and Technology,* 33(1), 93-101.

Dongare, A., Kharde, R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT),* 2(1), 189-194.

Dwivedi, G., & Sharma, M. (2014). Impact of cold flow properties of biodiesel on engine performance. *Renewable and Sustainable Energy Reviews,* 31, 650-656.

Dzulkefli, A. S., & Saad, N. M. (2020). The Performance of Energy Sector in Malaysia: Input-Output Analysis: An Extended Study. *Global Business & Management Research,* 12(4).

Emilio Martínez, Sonia Huertas, Héctor Ménez, José Luís Peñaa and Ana Alcaldea, (2008). Comparison of chemometric techniques applied to near infrared spectra for a gasoline blending control. *Journal of Near Infrared Spectroscopy*, volume 16: 297-303.

Esbensen, K. H., Guyot, D., Westad, F., & Houmoller, L. P. (2002). Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design. *Multivariate Data Analysis*

F Chauchard, J.M. Roger and V. Bellon-Maurel (2004). Correction of the temperature effect on Near Infrared Calibration – application to soluble solid content prediction. *J. Near Infrared Spectrosc,* 12 : 199-205

Faber, N., & Rajko, R. (2007). How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative. *Analytica Chimica Acta*, 595(1-2), 98-106.

Feng, F., Wu, Q., & Zeng, L. (2015). Rapid analysis of diesel fuel properties by near infrared reflectance spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,* 149, 271-278.

Florian Wulfert, Wim Th Kok and Age K. Smilde (1998). Influence of temperature on vibrational spectra and consequences for the prediction ability of multivariate models. *Analytical Chemistry* 70 (9) : 1761-1767.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109-135.

Gan, P. Y., & Li, Z. (2008). An econometric study on long-term energy outlook and the implications of renewable energy utilization in Malaysia. *Energy Policy,* 36(2), 890-899.

Gary, J. H., Handwerk, J. H., Kaiser, M. J., & Geddes, D. (2007). Petroleum refining: technology and economics: CRC press.

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

Ghosh, D., Moreira, J., and Mhaskar, P. (2022). Application of data-driven modeling approaches to industrial hydroprocessing units. *Chemical Engineering Research and Design* 177: 123-135.

Gonzaga, F. B., & Pasquini, C. (2010). A low cost short wave near infrared spectrophotometer: Application for determination of quality parameters of diesel fuel. *Analytica Chimica Acta,* 670(1-2), 92-97.

Gonzaga, F.B., and Pasquini, C. (2010). A low cost short wave near infrared spectrophotometer: Application for determination of quality parameters of diesel fuel.

Helga G. Aleme, Paulo J.S. Barbeira (2012): Determination of flash point and cetane index in diesel using distillation curves and multivariate calibration. *Fuel,* volume 102: 129-134.

Hoeil Chung, Hyuk-Jin Choi, Min-Sik Ku (1999). Rapid identification of petroleum products by near-infrared spectroscopy. *Korean Chem.Soc.,* volume 20: 1021-1025.

Hoeil Chung, Min-Sik Ku, Joon-Sik Lee (1999). Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene. *Vibrational Spectroscopy,* volume 20: 155-163.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417.

Hradecká, I., Velvarská, R., Jaklová, K.D., ands Vráblík, A. (2021). Rapid determination of diesel fuel properties by near-infrared spectroscopy. *Infrared Physics & Technology* 119: 103933.

Issam Barraa , Mourad Kharbacha,b , Mohamed Bousrabata , Yahia Cherraha , Mohamed Hanafic , El Mostafa Qannaric , Abdelaziz Bouklouzea (2019): Discrimination of diesel fuels marketed in Morocco using FTIR, *GC-MS analysis and chemometrics methods.* volume 209: 1-7.

Jianguo Sun, (1997). Statistical Analysis of NIR Data - Data Pretreatment. Journal of *Chemometrics*, volume 11: 525-532.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. Journal of the Royal Statistical Society: Series C (Applied Statistics), 31(3), 300-303.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical*, *Physical and Engineering Sciences,* 374(2065), 20150202.

Jones, D. S. (2008). Petroleum products and a refinery configuration. In Handbook of petroleum processing (pp. 47-109): Springer.

Julio Cesar L. Alves n , Ronei J. Poppi (2012). Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM). *Talanta,* volume 104: 155-161.

Kaisan, M., Anafi, F., Nuszkowski, J., Kulla, D., & Umaru, S. (2017). Calorific value, flash point and cetane number of biodiesel from cotton, jatropha and neem binary and multi-blends with diesel. *Biofuels*.

Kaixun He, Feng QIAN, Hui CHENG, and Wenli DU (2016). Improved integrated optimisation method of gasoline blend planning and real-time blend recipes. I & EC research: 1-42.

Knothe, G. (2010). Biodiesel: current trends and properties. *Topics in catalysis,* 53(11), 714-720.

Leonardo, R., Murta Valle, M., & Dweck, J. (2020). Thermovolumetric and thermogravimetric analysis of diesel S10. *Journal of Thermal Analysis and Calorimetry,* 139(2), 1507-1514.

Li, Z., Askim, J. R., & Suslick, K. S. (2018). The optoelectronic nose: colorimetric and fluorometric sensor arrays. Chemical reviews, 119(1), 231-292

Liu, S., Wang, S., Hu, C., Qin, *X*., Wang, J., & Kong, D. (2022). Development of a new NIR-machine learning approach for simultaneous detection of diesel various properties. *Measurement*, 187, 110293.

Loh, A, Soon, ZY, Ha, SY, and Yim, U.H. (2021). High-throughput screening of oil fingerprint using FT-IR coupled with chemometrics. *Science of The Total Environment* 760: 143354.

Loh, KH, (2016). Chemometric Approaches in the Evaluation of trace metals in commercially raised Tilapia and Preliminary Health Risk Assesments of its Consumption.

Long, J., Jiang, S., He, R., and Zhao, L. (2021). Diesel blending under property uncertainty: A data-driven robust optimisation approach. *Fuel* 306: 121647.

Low, K.H., Zain, S.M., Abas, M.R., Misran, M. and Mohd, M.A. (2009). Simultaneous spectrophotometric determination of copper, nickel, and zinc using 1-(2-thiazolylazo)-2-naphthol in the presence of triton x-100 using chemometric methods. *Journal of the Korean Chemical Society* 56: 717-726.

Malaysia Standard High PME Diesel Fuel Specification Euro 5, MS 123-5 (2020), Department of Malaysia Standard

Maleki, M.R., Mouazen, A.M., Ramon, H., and De Baerdemaeker, J. (2007). Multiplicative scatter correction during online measurement with near infrared spectroscopy. *Biosystems Engineering* 96(3): 427-433.

Máquina, A.D.V., Sitoe, V.B., Cruz, V.O., Santos, D.Q., and Neto, W.B. (2020). Analysis of 1H NMR spectra of diesel and crambe biodiesel mixtures using chemometrics tools to evaluate the authenticity of a Brazilian standard biodiesel blend. *Talanta* 109: 120590.

Martens, H., & Naes, T. (1984). Multivariate calibration. *Chemometrics* (pp. 147-156): Springer.

Martens, H., & Naes, T. (1989). Methods for calibration. *Multivariate calibration,* 1, 73-232.

Martens, H., & 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.

Martin, T. V., Zwally, H. J., Brenner, A. C., & Bindschadler, R. A. (1983). Analysis and retracking of continental ice sheet radar altimeter waveforms. *Journal of Geophysical Research: Oceans,* 88(C3), 1608-1616.

Melissa J. Romen, Michael J Adams, Andrew R. Hind, Suresh, K. Bhargava and Stephen. C. Grocott, (2002). Near-infrared prediction of oil yield from oil shale. *J. Near Infrared Spectrosc,* volume 10: 223-231.

Mertler, C. A., Vannatta, R. A., & LaVenia, K. N. (2021). Advanced and multivariate statistical methods: Practical application and interpretation: Routledge.

Miller, J., & Miller, J. C. (2018). Statistics and chemometrics for analytical chemistry: Pearson education.

Mishra, P., Biancolillo, A., Roger, J.M., Marini, F., and Rutledge, D.N. (2020). New data pre-processing trends based on ensemble of multiple pre-processing techniques. *TrAC Trends in Analytical Chemistry* 132: 116045.

Mishra, P., Marini, F., Biancolillo, A., and Roger J.M. (2021). Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques. *Talanta* 223: 121693.

Mochida, I., Okuma, O., & Yoon, S.-H. (2014). Chemicals from direct coal liquefaction. *Chemical reviews,* 114(3), 1637-1672.

Mohammadi, M., Khorrami, M.K., Vatani, A., Ghasemzadeh, H., Vatanparast, H., Bahramian, A., and Fallah, A. (2020). Rapid determination and classification of crude oils by ATR-FTIR spectroscopy and chemometric methods. Spectrochimica Acta Part A: *Molecular and Biomolecular Spectroscopy* 232: 118157.

Mohammadi, M.-R., Hadavimoghaddam, F., Atashrouz, S., Hemmati-Sarapardeh, A., Abedi, A., & Mohaddespour, A. (2022). Application of robust machine learning

methods to modeling hydrogen solubility in hydrocarbon fuels. *International Journal of Hydrogen Energy,* 47(1), 320-338.

Moro, M.K., dos Santos, F.D., Folli, G.S., Romão, w., and Filgueiras, P.R. (2021). A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy. *Fuel* 303: 121283.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). A user-friendly guide to multivariate calibration and classification (Vol. 6): NIR Chichester.

NM. Faber and R Rajikq (2007). How to avoid over-fitting in multivariate calibration – The conventional validation approach and an alternative. *Analytica Chimica Acta* 595 : 98-106.

Paiva, E.M., Ribessi, R.L., and Rohwedder, J.J.R. (2022). Near-infrared spectra of liquid and gas samples by diffuse reflectance employing benchtop and handheld spectrophotometers. *Spectrochimica Acta Part A: Molecular and Biomolecular* Spectroscopy 264: 120302.

Palou, A., Miró, A., Blanco, M., Larraz, R., Gómez, J. F., Martínez, T., . . . Alcalà, M. (2017). Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 180, 119-126.

Paolo Oliveri, Michele Fiorina (2012): Data analysis and chemometrics. *Chemical Analysis of Food :* 25-57.

Paricaud, P., Ndjaka, A., and Catoire, L. (2020). Prediction of the flash points of multicomponent systems: Applications to solvent blends, gasoline, diesel, biodiesels and jet fuels. *Fuel* 263: 116534.

Park, B., Chen, Y., Hruschka, W., Shackelford, S., & Koohmaraie, M. (1998). Near-infrared reflectance analysis for predicting beef longissimus tenderness. *Journal of animal science*, 76(8), 2115-2120.

Parkash, S. (2003). Refining processes handbook: Elsevier.

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives–A review. Analytica chimica acta, 1026, 8-36.

Peleg, Y. Shefer, S., Anavy, L., Chudnovsky, A., Israel, A., Golberg, A., and Yakhini, Z. (2019). Sparse NIR optimisation method (SNIRO) to quantify analyte composition with visible (VIS)/near infrared (NIR) spectroscopy (350 nm-2500 nm). *Analytica Chimica Acta 1051*: 32-40.

Peter de Pinder, Netherlands (2009): Characterisation and classification of crude oils using a combination of spectroscopy and chemometrics. Shell Global Solutions Nederland B.V. : 1-163.

Pirhadi, S., Shiri, F., & Ghasemi, J. B. (2015). Multivariate statistical analysis methods in QSAR. Rsc Advances, 5(127), 104635-104665.

Pontes, M. J. C., Pereira, C. F., Pimentel, M. F., Vasconcelos, F. V. C., & Silva, A. G. B. (2011). Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. *Talanta,* 85(4), 2159-2165.

Process Insights – Optical Absorption SpectroscopyJan 25 2019. (2023, March 10). Using near infrared spectrometery (NIR) to measure the cloud point of Diesel Fuel. Retrieved March 12, 2023, from https://www.azom.com/article.aspx?ArticleID=17514

Rana, M. S., Sámano, V., Ancheyta, J., & Diaz, J. (2007). A review of recent advances on process technologies for upgrading of heavy oils and residua. *Fuel,* 86(9), 1216-1231.

Retnam, A., Kassim, A. M., & Ahmad, W. K. W. (2015). Fingerprinting of light fuel oil: A Malaysia case study. *Procedia Environmental Sciences*, 30, 190-194.

Riazi, M. (2005). Characterization and properties of petroleum fractions (Vol. 50): ASTM international.

Roman M. Balabin, Ekaterina I. Lomakina, Ravilya Z. Safieva (2010). Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. *Fuel*, volume 90: 2007-2015.

Roman M. Balabin, Ravilya Z. Safieva (2008). Motor oil classification by base stock and viscosity based on near infrared (NIR) spectroscopy data. *Fuel,* volume 87: 2745-2752.

Roman M. Balabin, Ravilya Z. Safieva, Ekaterina I. Lomakina (2007). Comparison of linear and non-linear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction. *Chemometrics and intelligent laboratory system,* volume 88: 183-188.

Rosenfeld, H. J., Samuelsen, R. T., & Lea, P. (1998). Relationship between physical and chemical characteristics of carrots grown at northern latitudes. *The Journal of Horticultural Science and Biotechnology,* 73(2), 265-273.

Rossel, R.A.V., Walvoort, D.J.J., McBratney, A.B., Janik, I.J., and Skjemstad, J.O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma* 313: 59-75.

Santana, D. W. E. A., Sepulveda, M. P., & Barbeira, P. J. S. (2007). Spectrophotometric determination of the ASTM color of diesel oil. *Fuel,* 86(5-6), 911-914.

Santos Jr, V. O., Oliveira, F. C., Lima, D. G., Petry, A. C., Garcia, E., Suarez, P. A., & Rubim, J. C. (2005). A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis. *Analytica Chimica Acta,* 547(2), 188-196.

Santos, F.D., Santos, L.P., Cunha, P.H.P., Borghia, F.T., Romão, W., de Castro, E.V.R., de Oliveira, E.C., and Filgueiras, P.R. (2021). Discrimination of oils and fuels using a portable NIR spectrometer. *Fuel* 238: 118854.

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

Shafie, S. M., Mahlia, T. M. I., Masjuki, H. H., & Andriyana, A. (2011). Current energy usage and sustainable energy in Malaysia: A review. *Renewable and Sustainable Energy Reviews*, 15(9), 4370-4377.

Shenk, J., Westerhaus, M., & Templeton Jr, W. (1985). Calibration transfer between near infrared reflectance spectrophotometers 1. *Crop science,* 25(1), 159-161.

Shutao Wang, Shiyu Liu, Jingkun Zhang, Xiange Che, Zhifang Wang, Deming Kong (2019). Feasibility study on prediction of gasoline octane number using NIR spectroscopy combined with manifold learning and neural network. *Spectrchimica Acta part A: Molecular and Biomolecular Spectroscopy,* volume xx: 1-8.

Skrobot, V. L., Castro, E. V., Pereira, R. C., Pasa, V. M., & Fortes, I. C. (2007). Use of principal component analysis (PCA) and linear discriminant analysis (LDA) in gas chromatographic (GC) data in the investigation of gasoline adulteration. *Energy & fuels,* 21(6), 3394-3400.

Slinker, B., & Glantz, S. (1985). Multiple regression for physiological data analysis: the problem of multicollinearity. American Journal of Physiology-Regulatory, *Integrative and Comparative Physiology*, 249(1), R1-R12.

Smith, K., Willis, S., & Flinn, P. (1991). Measurement of the magnesium concentration in perennial ryegrass (Lolium perenne) using near infrared reflectance spectroscopy. *Australian journal of agricultural research,* 42(8), 1399-1404.

Stark, E., Luchter, K., & Margoshes, M. (1986). Near-infrared analysis (NIRA): A technology for quantitative and qualitative analysis. *Applied Spectroscopy Reviews*, 22(4), 335-399.

Stuntz, G., & Plantenga, F. (2002). New technologies to meet the low sulfur fuel challenge. Paper presented at the 17th World Petroleum Congress.

Svante Wold, Michael Sjostrom, Lennart Eriksson, (2001). PLS- regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58: 109–130.

Swarup, S., & Schoff, C. K. (1993). A survey of surfactants in coatings technology. *Progress in organic coatings*, 23(1), 1-22.

Taavitsainen, V.-M. (2009). Denoising and signal-to-noise ratio enhancement: derivatives.

Thomas, A. (1988). Automotive fuels. In Internal combustion engines (pp. 213-270): Elsevier.

Ulmschneider, M., & Roggo, Y. (2008). Process analytical technology. Gad SC (Ed), Pharmaceutical Manufacturing Handbook, 353-410.

Vandeginste, B. M., Massart, D., Buydens, L., De Jong, S., Lewi, P., & Verbeke, J. (1998). Handbook of Chemometrics and Qualimetrics.

Verma, A., Nimana, B., Olateju, B., Rahman, M. M., Radpour, S., Canter, C., Kumar, A. (2017). A techno-economic assessment of bitumen and synthetic crude oil transport (SCO) in the Canadian oil sands industry: Oil via rail or pipeline? *Energy,* 124, 665-683.

Vinicius L. Skrobot, Eustaquio V.R. Castro, Rita C. C. Pereira, Vanya M. D. pasa, Isabel C.P. Fortes. (2007). Use of principal component analysis and linear discriminant analysis in gas chromatographic data in the investigation of gasoline adulteration. *Energy and Fuel* 21: 3394-3400.

Wang, S., Liu, S., Yuan, Y., Zhang, J., Wang, J., & Kong, D. (2020). Simultaneous detection of different properties of diesel fuel by near infrared spectroscopy and chemometrics. *Infrared Physics & Technology,* 104, 103111.

Wang, S., Liu, S., Yuan, Y., Zhang, J., Wang, Z., and Che, *X*. (2020b). A novel CC-tSNE-SVR model for rapid determination of diesel fuel quality by near infrared spectroscopy. *Infrared Physics & Technology* 106: 103276.

Wei Zhang, Hang Song, Jing Lu, Wen Liu, Lirong Nie, and Shun Yao (2015). Online nir analysis and prediction model for synthesis process of ethyl 2-chloropropionate. *International Journal of Analytical Chemistry,* volume 2015: 1-7.

Weyer, L. G. (1985). Near-infrared spectroscopy of organic substances. *Applied Spectroscopy Reviews,* 21(1-2), 1-43.

Williams, P. (1987). Variables affecting near-infrared reflectance spectroscopic analysis. Near-infrared technology in the agricultural and food industries, 143-167.

Williams, P., Burns, D., & Ciurczak, E. (1992). Handbook of Near-Infrared Analysis. Burns, DA.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems, 58(2), 109-130.

Xiaoli Wang, Guomin Zhou, (2017). Study of pre-treatment algorithm of near infrared spectroscopy. HAL: 623-632.

Yahya, S.I., and Aghel, B. (2021). Estimation of kinematic viscosity of biodiesel-diesel blends: Comparison among accuracy of intelligent and empirical paradigms. *Renewable Energy* 177: 318-326.

Zhanfeng Xu, Christopher E. Bunker, and Peter De B. Harrington (2010). Classification of jet fuel properties by near-infrared spectroscopy using fuzzy rule-building expert systems and support vector machines. *Society for Applied Spectroscopy, volume* 64, Number 11: 1251-1258.

Zhang, *X*., Li, H., Zhang, Y., Qi, H., Yang, *X*., Wang, Q., and Li, D. (2021). Quantitative analysis of the oil mixture using PLS combined with spectroscopy detection. *Optik* 244: 167611.

Zvirin, Y., Gutman, M., & Tartakovsky, L. (1998). Fuel Effects on Emissions. Chapter 14, Handbook of Air Pollution from Internal Combustion Engines: Pollutants Formation and Control, edited by E. Sher. In: Academic Press.