Elsevier Editorial System(tm) for Fuel
Manuscript Draft

Manuscript Number: JFUE-D-16-04041R3

Title: Detection and Estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with Multivariate Methods

Article Type: Research paper

Keywords: NIR- spectroscopy; gasoline, adulteration; PCA, PLS-DA, PLS regression

Corresponding Author: Dr. Fazal Mabood, Ph.D

Corresponding Author's Institution: University of Nizwa

First Author: Fazal Mabood, Ph.D

Order of Authors: Fazal Mabood, Ph.D; Farah Jabeen, PhD; Syed A Gillani, PhD; Javid Hussain, PhD; Ahmed Hamaed, PhD

Abstract: Super premium 95 octane gasoline is a special blend of petrol with a higher octane rating that can produce higher engine power, as well as knock-free performance for cars with a high-octane requirement. Super premium grade gasoline 95 is often adulterated with cheaper Premium grade 91 that lowers the octane number of the Super premium gasoline. In the present study a new Near Infrared (NIR) spectroscopy combined with multivariate analysis was developed to detect as well as to quantify the level of Premium 91 gasoline adulteration in Super premium 95 octane gasolines. In this study standard samples of Premium 91 and Super premium 95 octane gasoline were collected from Oman Oil Refineries and Petroleum Industries Company SAOC (ORPIC) and were investigated. Super premium 95 samples were then adulterated with eighteen different percentage levels: 0%, 1%, 3%, 5%, 7%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, and 75% of Premium 91 gasoline. All samples were measured using NIR spectroscopy in absorption mode in the wavelength range from 700-2500 nm. The multivariate methods like PCA, PLS-DA and PLS regression were applied for statistical analysis of the obtained NIR spectral data. Partial least-squares discriminant analysis (PLSDA) was used to check the discrimination between the pure and adulterated gasoline samples. For PLSDA model the R-square value obtained was 0.9984 with 0.0198 RMSE. Furthermore, PLS regression model was also built to quantify the levels of Premium 91 adulterant in Super Premium 95 gasoline samples. The PLS regression model was obtained with the R-square 0.99 and with 1.90 RMSECV value having good prediction with RMSEP value 1.98 and correlation of 0.99. This newly developed method is having lower limit of detection less than 2 % level for Premium 91 adulteration. It was desirable to have simple, rapid and sensitive methods to detect the presence of one petroleum product in another.

Dr. Zuohua Huang

Principal Editor,

Fuel

Subject: Resubmission of Manuscript entitled, Detection and Estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with Multivariate Methods

Dear Dr. Huang,

Thank you very much for your quick and efficient action on our submission. It is highly appreciated. Enclosed please find the revised version of the subject manuscript which has been modified as per the reviewer suggestions.

I hope that it would be suitable for consideration in your esteemed Journal.

Looking forward for a positive response

Regards

Fazal Mabood (Ph.D),

Associate Professor of Analytical Chemistry,
Head of Chemistry Section (HoS),
Department of Biological Sciences & Chemistry
College of Arts and Sciences,
University of Nizwa,Sultanate of Oman
Sultanate of OmanTel. 0096895971085
e-mail: mehboob@unizwa.edu.om
e-mail; mehboob86@yahoo.com
URL:www.unizwa.edu.om

Dear Editor

Thank you very much your kind suggestions on the manuscript Ref: JFUE-D-16-04041R1 "**Detection and Estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with Multivariate Methods**" submitted to Fuel. To incorporate the suggestions of reviewers in the manuscript made a big difference. I cordially appreciate it, thank you very much for their precious time.

Remarks of Reviewer #1

Reviewer #1: Major amendments
Remarks: There is no doubt that detecting fraud is very important, both in developing countries and in developed ones. My concern about this paper was not on this topic.
As a chemometrics practitioner, I do not like papers that use a black box to calculate something. This paper uses PLS and PLSDA as such, without giving attention to the reason or the fundamentals. I know that the first derivative shows differences but it also introduces numerical modifications, specially using smoothing like the Savitzky-Golay algorithms do. So, the use of the first derivatives instead of the original or MSC-corrected spectra may not be adequate and results should be provided for both cases to justify the choice.
In the same way, the black box does not show what is going on. What are the significant wavelengths and what do they represent? These are fundamental questions in a research paper and are not addressed here. A simple representation of the loadings gives an insight on the basics and should accompany every PLS paper.
All of the above causes me a profound displeasure with the paper.In my first review I indicated this but there have been little changes in the paper. There have been other attempts to similar problems and this paper should go farther to be published. Unfortunately, that is not true in its actual state. I cannot recommend its publication..

Reply: Dear Reviewer I am cordially thankful for your highly valuable suggestions. Believe me I really appreciate it and it has improved the quality of the manuscript, thank you

Point1. To justify the choice of pre-treatment I have tabulated all the parameters in Table 1 as below.

Table 1. Selection of Pre-processing

Type of spectra	Pre-processing	PLS		PLSDA		PLS		# of factors
		RMSEC	R^2	RMSEC	R^2	RMSEP	R^2	
Full Spectra (4000 to 10000 cm^{-1})	Without pre-processing	2.406	0.99	0.013	0.99	1.87	0.99	4
Full Spectra (4000 to 10000 cm^{-1})	MSC	2.406	0.99	0.318	0.75	9.64	0.82	4
Spectra (4000 to	Without	1.63	0.99	0.164	0.88	1.97	0.99	4

6500 cm^{-1})	pre-processing							
Spectra (4000 to 6500 cm^{-1})	MSC	1.76	0.99	0.165	0.88	1.56	0.99	4
Full Spectra (4000 to 10000 cm^{-1})	1^{st} derv. with 11 smoothing points	1.59	0.99	0.045	0.99	1.65	0.99	3
Spectra (4000 to 6500 cm^{-1})	1^{st} derv. with 11 smoothing points	1.33	0.99	0.012	0.99	1.35	0.99	3

As it can be seen from Table 1 that the application of 1^{st} derivative functions with Savitzky-Golay smoothing 11 points for wavelength range 4000 to 6500 cm^{-1} has improved the parameters like RMSEC, R^2 as well as RMSEP. All of them have minimum error and the highest correlationship with less number of factors used i.e. 3 factors for both PLS as well as PLSDA models. All the remaining models were built by using the 1^{st} derivative functions with Savitzky-Golay smoothing 11 points for wavelength range 4000 to 6500 cm^{-1} at 2 polynomial orders .

Point2. Representation of the loadings gives

Similary the factor loading plot that is analogous to correlation coefficients, by squaring them give the amount of explained variation for PLS model is shown in Figure 7.

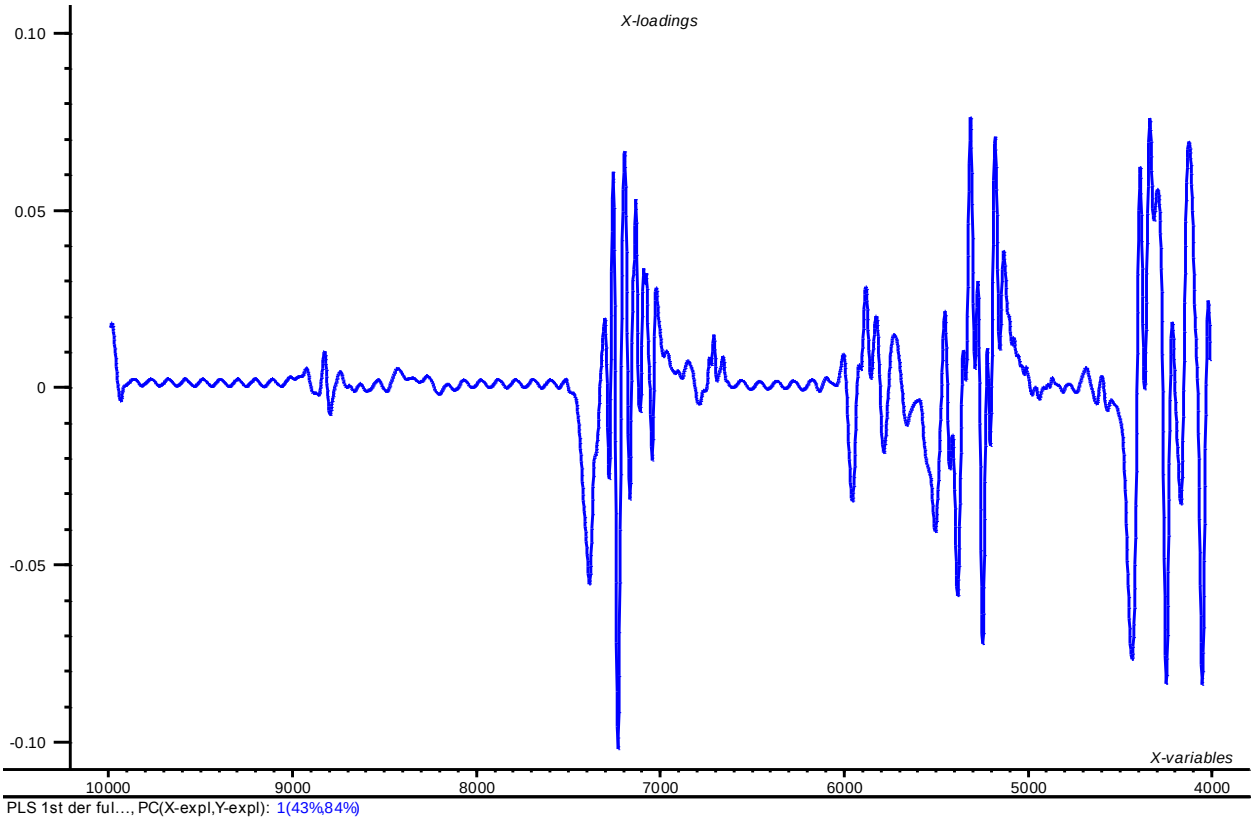


Figure 7a. Factor loading plot for factor 1

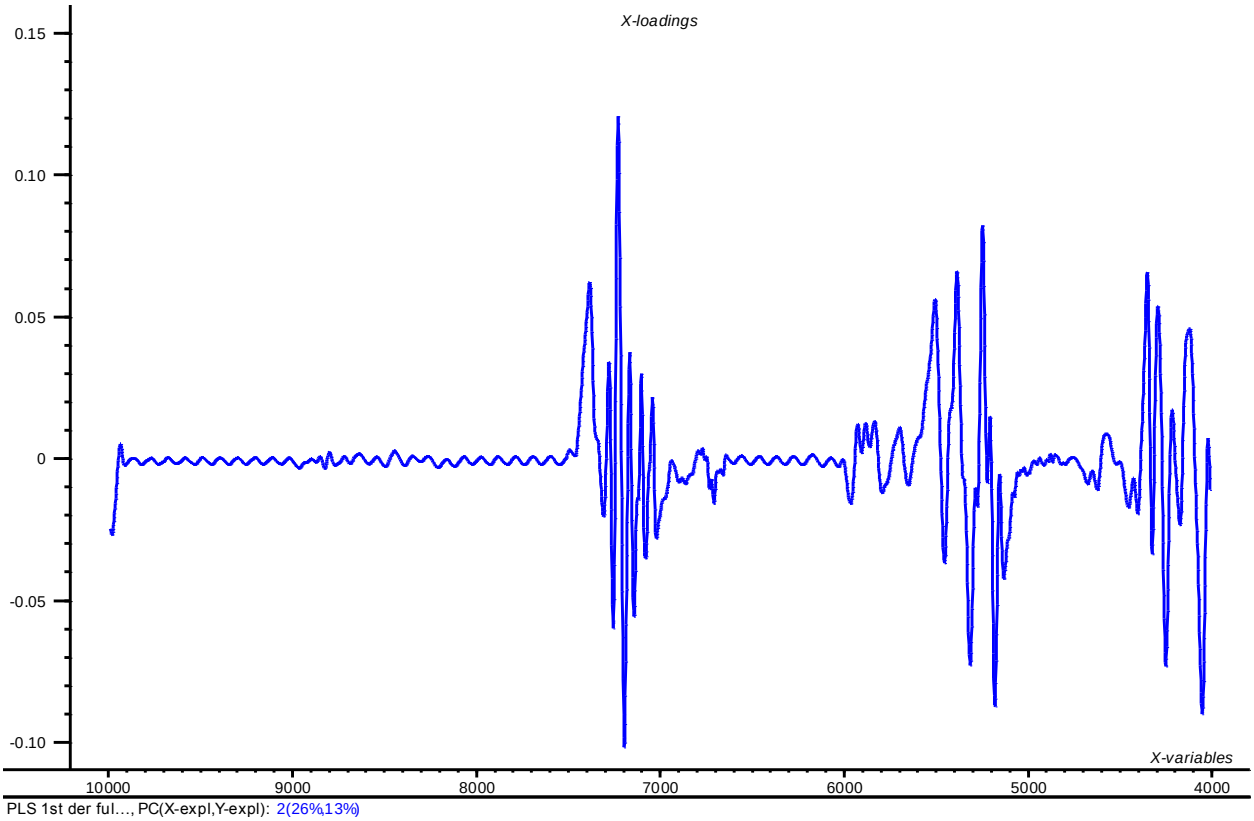


Figure 7b. Factor loading plot for factor 2

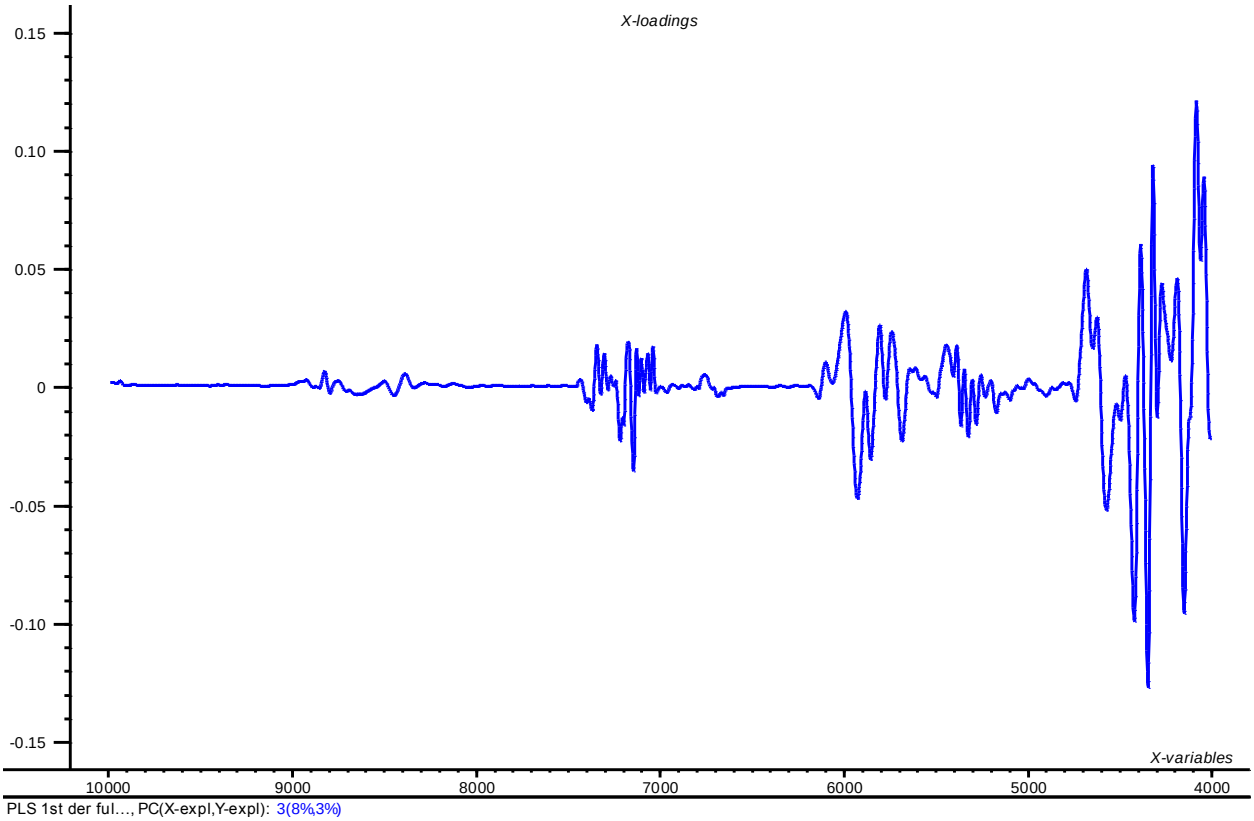


Figure 7c. Factor loading plot for factor 3

Figure 7a,b and c shows the factor loading plot for factor 1. It tell us how much of the variation in a variable is explained by the factor. In this case, 3 factors contain 77% of the total variation. Factor 1 explains 43 % of the variation, factor 2 explains 26%, and factor 3 explains 8%. The remaining 3 components explain only 21%.

All of the suggested changes have been incorporated in the manuscript. I am also going to submit the additional supporting file for other models and figures.

I must say many many thanks

Regards

Fazal

O.K.

Highlights

- Development of New NIR spectroscopy with multivariate methods for detection & estimation of gasoline adulteration
- To build PLSDA, PCA models as detection & exploration tools
- To build PLS regression model as quantification tool.

Detection and Estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with Multivariate Methods

Fazal Mabood*[a], Syed Abdullah Gilani*[a], Mohammed Albroumi[a], Saif Alameri[a], Mahmood M. O. Al Nabhani[a], Farah Jabeen*[b], Javid Hussain[a], Ahmed Al-Harrasi[e], Ricard Boqué[c], Saima Farooq[a], Ahmad M.Hamaed[a], Zakira Naureen[a], Alamgir Khan[f] and Zahid Hussain[d]

a) *Department of Biological Sciences & Chemistry, College of Arts and Sciences, University of Nizwa, Sultanate of Oman, (mehboob@unizwa.edu.om,. gilani@unizwa.edu.om, fjabeen2009@yahoo.com)*

b) *Department of Chemistry, University of Malakand, KPK, Pakistan.*

c) *Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, Tarragona, Spain*

d) *Department of Chemistry, Abdul Wali Khan University, KPK, Pakistan*

e) *UoN Chair of Oman Medicinal Plants and Marine Products, University of Nizwa, Sultanate of Oman.*

f) *Department of Chemistry and Biology (DQB/CECEN),State University of Maranhão (UEMA) São Luis - MA, Brazil*

Abstract

Super premium 95 octane gasoline is a special blend of petrol with a higher octane rating that can produce higher engine power, as well as knock-free performance for cars with a high-octane requirement. Super premium grade gasoline 95 is often adulterated with cheaper Premium grade 91 that lowers the octane number of the Super premium gasoline. In the present study a new Near Infrared (NIR) spectroscopy combined with multivariate analysis was developed to detect as well as to quantify the level of Premium 91 gasoline adulteration in Super premium 95 octane gasolines. In this study standard samples of Premium 91 and Super premium 95 octane gasoline were collected from Oman Oil Refineries and Petroleum Industries Company SAOC (ORPIC) and were investigated. Super premium 95 samples were then adulterated with eighteen different percentage levels: 0%, 1%, 3%, 5%, 7%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%,

55%, 60%, 65%, 70%, and 75% of Premium 91 gasoline. All samples were measured using NIR spectroscopy in absorption mode in the wavelength range from 700-2500 nm. The multivariate methods like PCA, PLSDA and PLS regression were applied for statistical analysis of the obtained NIR spectral data. Partial least-squares discriminant analysis (PLSDA) was used to check the discrimination between the pure and adulterated gasoline samples. For PLSDA model the R-square value obtained was 0.99 with 0.012 RMSE. Furthermore, PLS regression model was also built to quantify the levels of Premium 91 adulterant in Super Premium 95 gasoline samples. The PLS regression model was obtained with the R-square 0.99 and with 1.33 RMSECV value having good prediction with RMSEP value 1.35 and correlation of 0.99. This newly developed method is having lower limit of detection less than 1.5% level for Premium 91 adulteration. It was desirable to have simple, rapid and sensitive methods to detect the presence of one petroleum product in another.

Keywords

NIR- spectroscopy; gasoline, adulteration; PCA, PLS-DA, PLS regression

1. Introduction

Fuels, on which the world's industries, economies, and daily lives depend, have become a crucial part of life of every human being. The governments of every country due to current geopolitical situations, wars and fluctuating economies have imposed heavy taxations on fuels that hiked their prices abnormally, for example, in south Asia, higher taxes are imposed on gasoline followed by diesel, kerosene, industrial solvents and recycled lubricants. Due to heavy taxes, especially differential taxing system, adulterations in the fuels are common practices [1, 2]. Adulteration of fuels is to mix expensive product i.e., super premium gasoline with the cheaper product i.e., regular grade gasoline or mixing of diesel fuel with cheaper light heating oil. Detection of gasoline adulteration, especially when it is with lower percentage (10 to 30% by volume) cannot be easily done [2]. Therefore, in some of the countries, for example, in India, illegal selling of adulterated gasoline mixed with diesel, and diesel mixed with kerosene is in common practice [2]. From the financial point of view, less than 10% adulteration is not much beneficial for dealers or sellers while more than 30% adulteration would cause decreasing engine performance of the vehicle and can be detected [2].

Whenever the combustion quality of the gasoline and anti-knock quality or resistance to pre-ignition is determined, an average of Research Octane Number (RON) is used [3,4]. One of the expensive and high octane gasoline is Super premium 95 that is the blending of petrol with a higher octane rating 96. The consumers prefer Super premium 96 for higher engine power, knock-free performance. In the market, adulteration of Super premium gasoline grade 96 with cheaper premium grade 91 results in lowering down the octane number of the Super premium gasoline.

To detect adulteration in gasoline, several methods of combining chemometric tools with conventional techniques of gasoline analysis have been used [5 – 16]. Most of these methods are based on chromatographic and spectroscopic studies [5 – 16]. Balabin and Safieva (2008) used near infrared (NIC) spectroscopy method with three different analytical methods i.e., linear discriminant analysis (LDA), soft independent modeling of class analogy (SIMCA), and multilayer perceptron (MLP) and classified 382 gasoline samples and fractions [17]. They reported that NIR spectroscopy along with MPLP technique was effective method for classification.

Nine different multivariate classification methods such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized discriminant analysis (RDA), soft independent modeling of class analogy (SIMCA), partial least squares (PLS) classification, K-nearest neighbor (KNN), support vector machines (SVM), probabilistic neural network (PNN), and multilayer perceptron (ANN-MLP) for gasoline classification showed that KNN, SVM, and PNN techniques for classification were found to be among the most effective ones [18]. However, poor results were observed by using Artificial neural network (ANN-MLP) approach based on principal component analysis (PCA).

The quality of Brazilian gasoline was studied in 47 commercial samples and 21 intentionally adulterated samples with organic solvents using 1H NMR (Nuclear Magnetic Reasonance) spectroscopy [16]. Chemometric methods such as Principal Component Analysis and Heirarchical Cluster Analysis were applied. The results grouped commercial samples into conform and adulterated ones into the *nonconform* groups with the tendency of increasing solvent concentration [16]. In another study of identifying heavy aliphatic, light aliphatic and aromatic hydrocarbons in Brazilian gasoline samples, studies of physicochemical properties were

combined with GC (gas chromatographic) analysis followed by multivariate analysis showed significant results that could differentiate between adulterated and non-adulterated samples [5].

At industrial level, NIR spectroscopic methods have been used for extraction and quantification of different products, such as crude extracts or pure compounds that must have direct or indirect absorbance. NIR spectra can be retrieved within short time period but to find correlation between spectral characteristics and the properties needs data analysis and modeling phase that is time consuming process. To build the chemometric models on training-set samples, the databases are prepared, based on spectral absorbencies and correlated reference labs, and applied [5-10]. NIR spectroscopy has significantly determined the quality of gasoline on the basis of octane number, ethanol contents, MTBE (nethyl tert-butyl ether) content, distillation points, Reid vapor pressure, aromatic and saturated contents [5 – 10]. NIR spectroscopy has shown far better results than using gas chromatography or Nuclear Magnetic Resonance (NMR) [18].

The spectrum may be interpreted qualitatively and quantitatively when physic-chemical properties are combined with chemometric methods such as Multivariate regression procedures. Partial Least Squares (PLs) method in multivariate regression analysis is used to establish a relationship between physico-chemical properties (dependent Y variable) and measured spectra for all samples (independent X variable). In case of multiple variables, when considering unknown gasoline samples from the same refinery even, univariate analysis will produce false results, therefore, multivariate regression models must be built [19].

Adulteration may be effectively studied, if the fuel quality is checked at distribution points with portable inexpensive equipment, quick measurement methods, and quick results on the spot. For this purpose, adulteration in Super premium gasoline was studied. To identify adulterations in Super premium gasoline, NIR spectroscopy method was combined with chemometric techniques of classification PCA and PLS-DA. Furthermore, multivariate calibration models were built using PLS of prediction of the premium 91 adulteration in super premium 96 gasoline.

2. Experimental

2.1. Adulterated Samples preparation

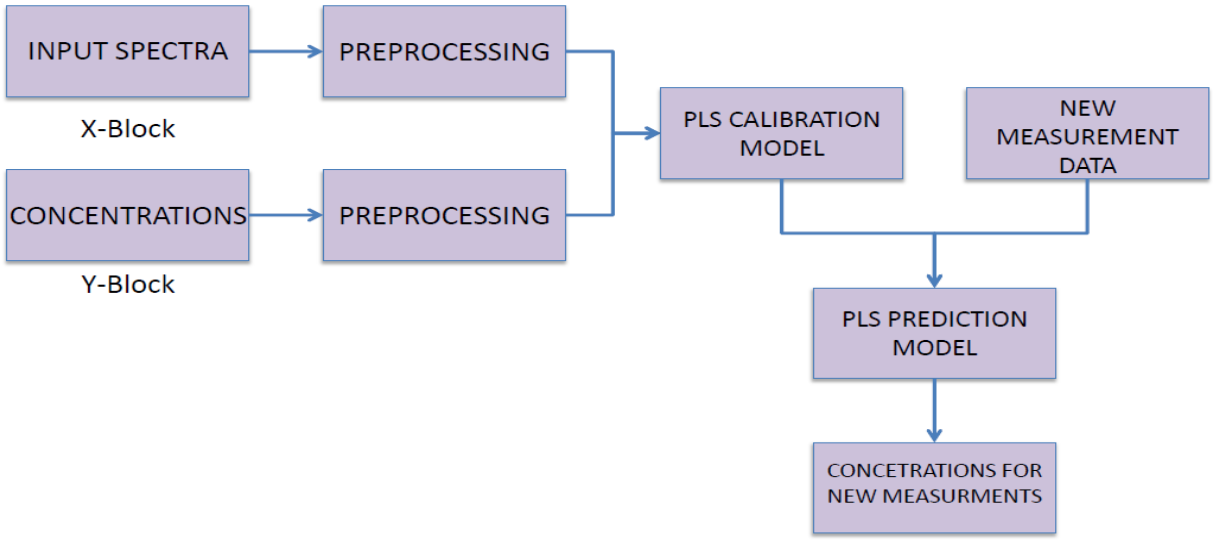
In this study standard samples of Premium 91 and Super premium 95 octane gasoline were collected from Oman Oil Refineries and Petroleum Industries Company SAOC (ORPIC). Super premium 95 samples were then adulterated with eighteen different percentage levels: 0%, 1%, 3%, 5%, 7%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, and 75% of Premium 91 gasoline each in triplicate. The total number of samples used was 57. For PLS regression all the samples were joined together and split into two sets, a training set (70% of the samples) and a test set for validation (30% of the samples).

2.2. NIR Spectroscopic analysis

All samples were measured using the Frontier™ IR/NIR system model number (L1280034) by PerkinElmer in absorption mode in the wavelength range from 700-2500 nm, at 2 cm^{-1} resolution and using a 0.2 mm path length CaF_2 sealed cell. Prominent absorption peaks were appeared in the region from 4000 to 7588 cm^{-1} wavenumber.

2.3. Statistical analysis

For statistical analysis, Unscrambler version 9.0 and Microsoft Excel 2010 were used. PCA, PLSDA and PLS models were applied on both pure and adulterated gasoline samples. Multivariate calibration technique such as partial-least squares regression was used to construct a mathematical model that relates the multivariate response (spectrum) to the concentration of the analyte of interest, and such a model can be used to efficiently predict the concentrations of new samples. The use of rank reduction techniques such as discriminant analysis on principal components or partial least squares scores. Spectral pretreatments such as SNV, baseline correction and S.Golay smoothing of 13 points were applied. Full cross validation was used for building PLS-DA model. External cross validation was used to validate the PLS regression models built with the training set. RMSECV (Root mean square error of cross validation) was used as an internal indicator of the predictive ability of the models.



142

143 Figure A. Flowsheet diagram of PLS regression analysis

144

145 **3. Results and discussion**

146 *3.1. Near Infrared spectra*

147 Figure 1 shows the NIR spectra of all the samples ranging from 10000-4000 cm^{-1} in term of
148 wavenumbers while in term of wavelength ranging from 700-2500 nm using a 0.2mm path
149 length CaF_2 sealed cell.

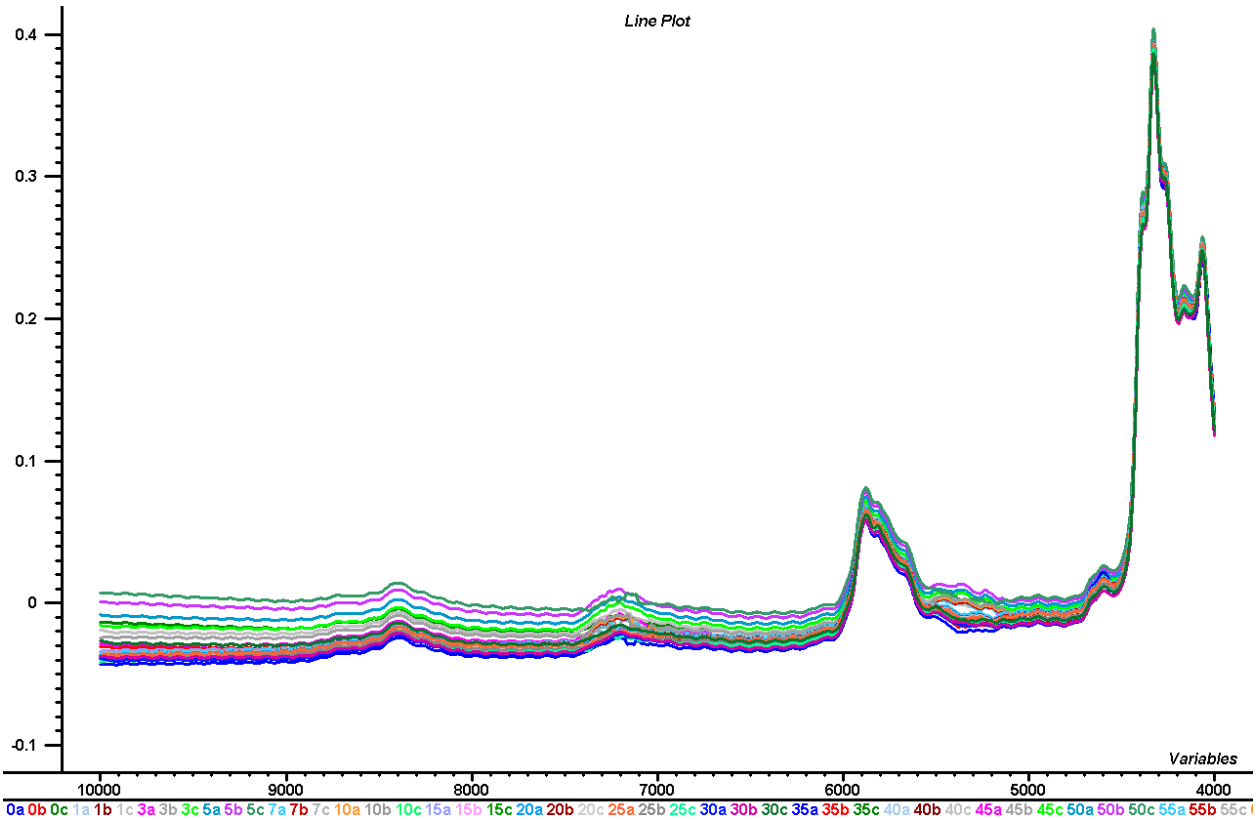


Figure 1. NIR spectra of premium 91 and super premium 95 octane gasoline samples without pre-processing

The spectra in Figure 1 shows a scattering effect due to reflection and it is also not very smooth. Various types of spectral pretreatments, such as MSC, 1st derivative as shown in Table 1 were applied.

Table 1. Selection of Pre-processing

Type of spectra	**Pre-processing**	**PLS**		**PLSDA**		**PLS**		**# of factors**
		RMSEC	**R^2**	**RMSEC**	**R^2**	**RMSEP**	**R^2**	
Full Spectra (4000 to 10000 cm^{-1})	Without pre-processing	2.406	0.99	0.013	0.99	1.87	0.99	4
Full Spectra (4000 to 10000 cm^{-1})	MSC	2.406	0.99	0.318	0.75	9.64	0.82	4
Spectra (4000 to 6500 cm^{-1})	Without pre-	1.63	0.99	0.164	0.88	1.97	0.99	4

	processing							
Spectra (4000 to 6500 cm^{-1})	MSC	1.76	0.99	0.165	0.88	1.56	0.99	4
Full Spectra (4000 to 10000 cm^{-1})	1^{st} derv. with 11 smoothing points	1.59	0.99	0.045	0.99	1.65	0.99	3
Spectra (4000 to 6500 cm^{-1})	**1^{st} derv. with 11 smoothing points**	**1.33**	**0.99**	**0.012**	**0.99**	**1.35**	**0.99**	**3**

As it can be seen from Table 1 that the application of 1^{st} derivative functions with Savitzky-Golay smoothing 11 points for wavelength range 4000 to 6500 cm^{-1} has improved the parameters like RMSEC, R^2 as well as RMSEP. All of them have minimum error and the highest correlationship with less number of factors used i.e. 3 factors for both PLS as well as PLSDA models. All the remaining models were built by using the 1^{st} derivative functions with Savitzky-Golay smoothing 11 points for wavelength range 4000 to 6500 cm^{-1} at 2 polynomial order as shown in Figure 2. It also shows the noisy region in between 7000 cm^{-1} to 7500 cm^{-1}.

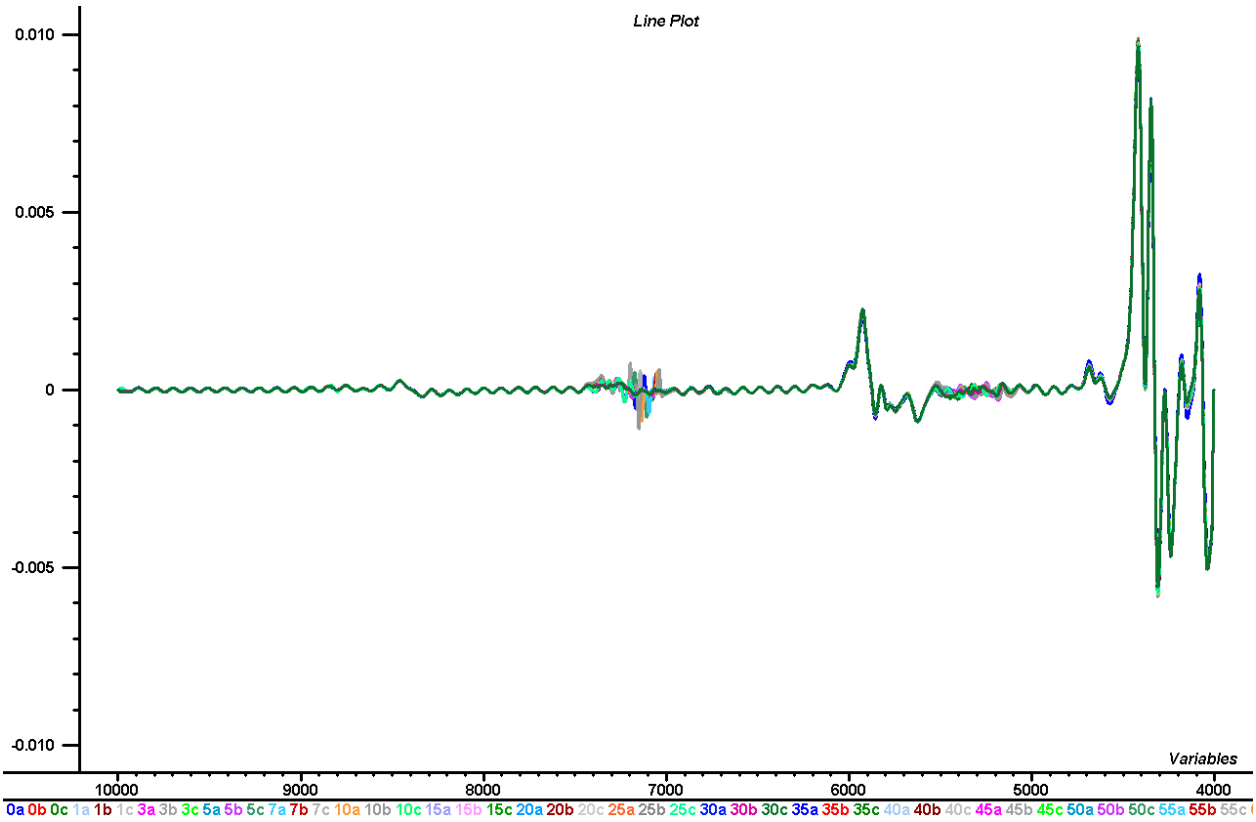


Figure 2. NIR spectra after 1st derivative functions with Savitzky-Golay smoothing 11 points for wavelength range 4000 to 6500 cm^{-1} at 2 polynomial orders

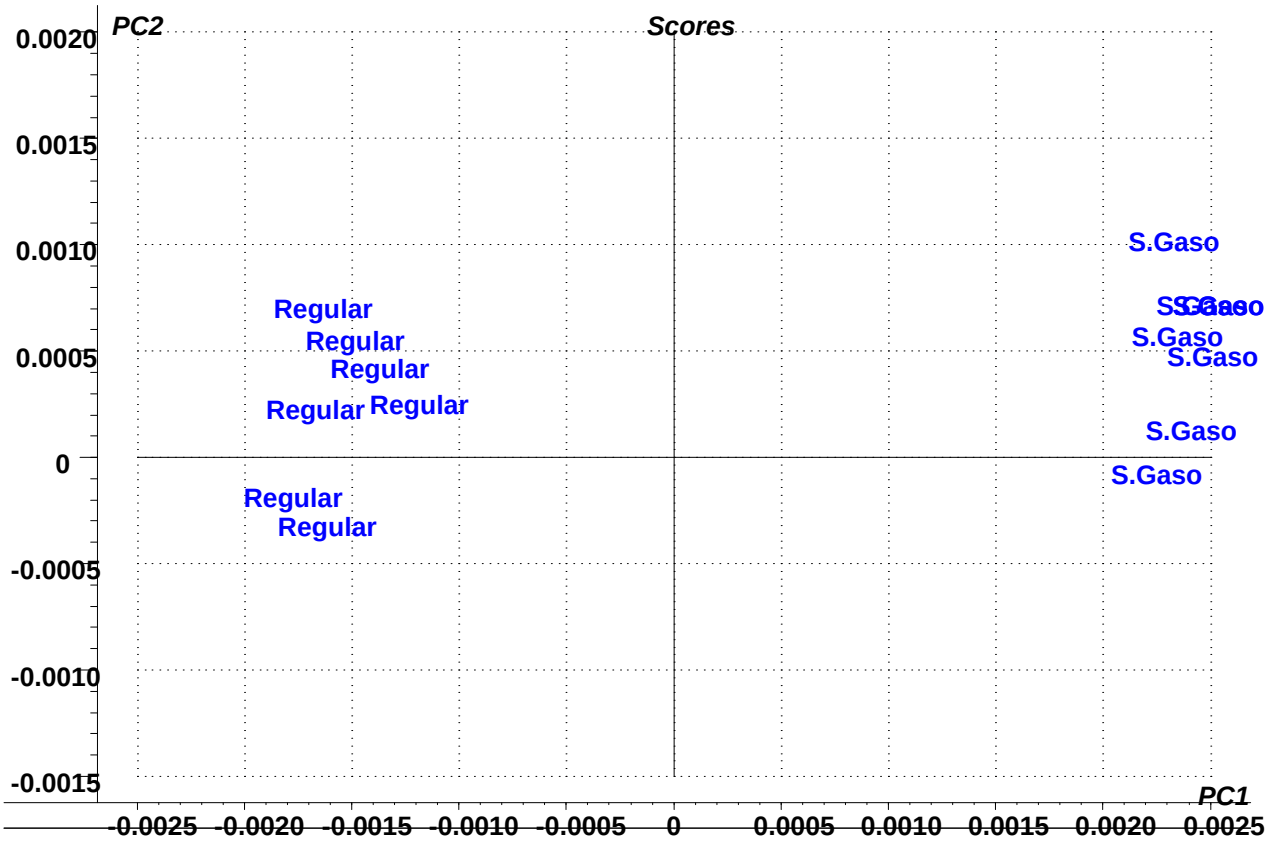
It can be seen from the spectra in figures 2 that there are prominent absorption peaks in between wavenumber 4000 cm^{-1} to 6500 cm^{-1} for all the samples.

In order to visualize more the effect of variation between the premium 91 and super premium 95 octane gasoline samples an alternative approach of principal components analysis (PCA), was applied. PCA model was built as shown in Figure 3. PCA is a standard multivariate data analysis exploratory tool. It is used to reduce the dimensionality of a complex data set without much loss of information, to extract the most important information from the data table, to identify noise and outlier in the data set. It is a way of identifying the underlying patterns in data for further analysis using other techniques. The procedure of PCA is like that it converts a set of correlated variables into a new set of uncorrelated variables called principal components. PCA redistributes the total variance of the data set in such a way that the first principal component has maximum variance, followed by second component and so on.

Variance PC1 > Variance PC2 > … Variance PCk

Total variance = Variance PC1 + Variance PC2 + … Variance PCk

The covariance of any of the principal component with any other principal component is zero (uncorrelated) and they are orthogonal to each other.



RESULT4, X-expl: 80%,19%

Figure 3. PCA score plot of pure premium 91 and super premium 95 octane gasoline samples

It can be seen from the PCA score plot that there is complete classification and separation in between premium 91 and super premium 95 octane gasoline samples. They are spaced and grouped in the specific different regions of the PCA score plot.

Similarly, PLSDA model was also built for spectral data between pure super premiums 95 and with the samples adulterated with 10% of premium 91 gasoline as shown in Figure 5. PLS Discriminant Analysis (PLSDA) is performed in order to sharpen the separation between groups of observations, by rotating the PCA components such that a maximum separation among classes

is obtained, and to understand which variables carry the class separating information. PLSDA model can be used as an identifaction tool to check premium 91 adulteration in super premium 95 octane gasoline samples. If there is any amount of premium 91 adulteration they will occupy the space in between the pure and adulterated samples as shown in Figure 4.

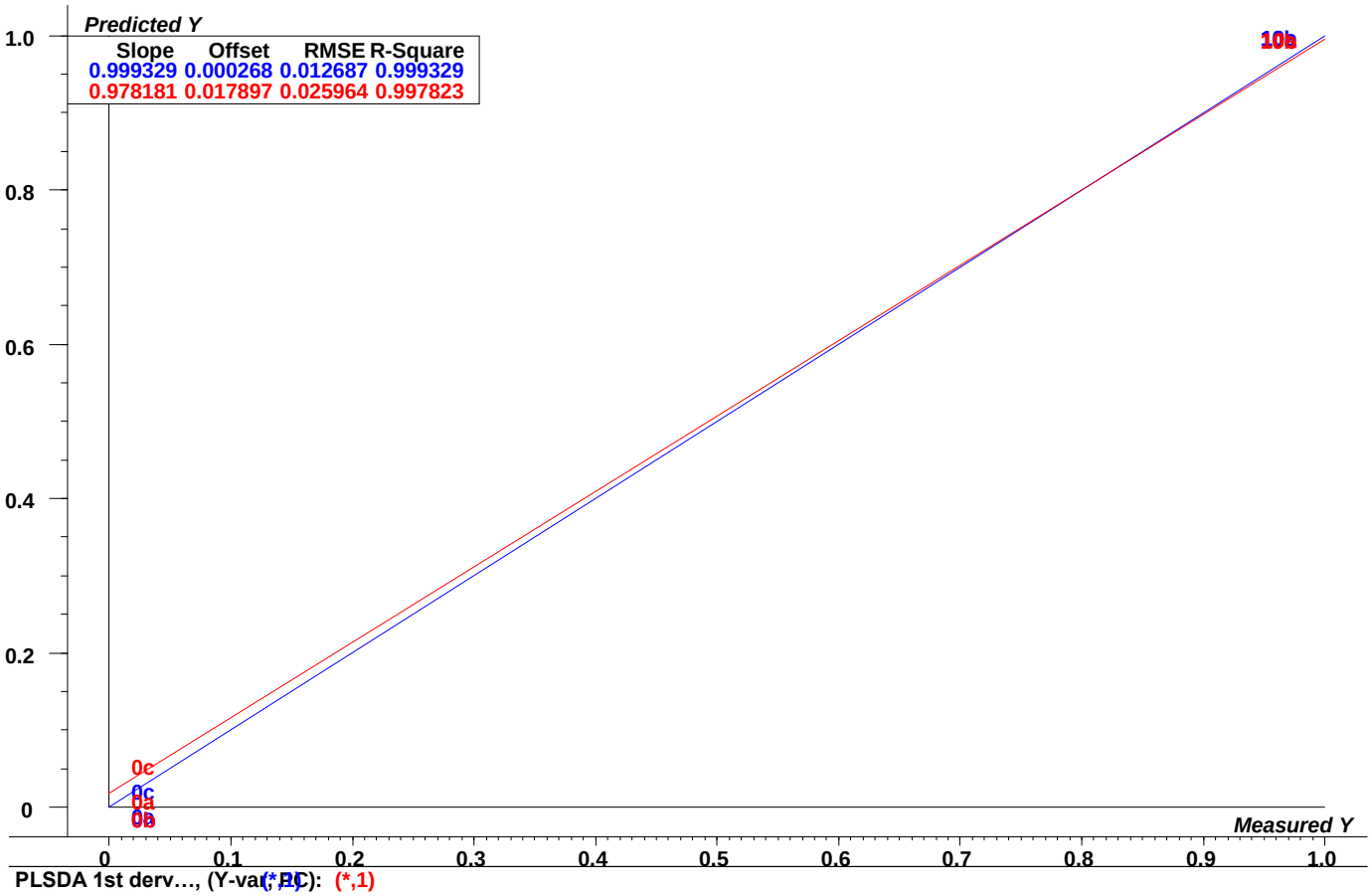


Figure 4. PLS-DA model after 1st derivative spectral treatment in the wavenumber range (6500 to 4000 cm^{-1}) for pure super premium 95 and with 10 % premium 91 adulteration

It can be seen from Figure 4 that pure Super premium 95 samples are completely discriminated from 10 % Premium 91 adulterated samples. If some sample is having adulteration that will be appearing in the middle region. This is the reason PLSDA model is used as an adulteration detection tool. The RMSECV value for PLSDA model was found 0.0126with with R square value of 0.99.

3.2 PLS regression results

To quantify the level of premium 91 adulteration in super premium 95 gasoline samples PLS regression model was built by using 70 % of the samples as a training set with eighteen different percentage levels: 0%, 1%, 3%, 5%, 7%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, and 75% of premium 91 gasoline.. PLS regression models are shown in Figure 5. PLS is a predictive technique and it is particularly useful when predictor variables are highly correlated or when the number of predictors exceeds the number of cases. PLS combines features of principal components analysis and multiple regression. It first extracts a set of latent factors that explain as much of the covariance as possible between the independent **X** and dependent **Y** variables. Then a regression step predicts values of the dependent variables using the decomposition of the independent variables. PLS finds a set of orthogonal components that maximize the level of explanation of both X and Y provides a predictive equation for Y in terms of the X's. PLSR derives its usefulness from its ability to analyze data with many, noisy, collinear, and even incomplete variables in both **X** and **Y**. PLS has the desirable property that the precision of the model parameters improves with the increasing number of relevant variables and observations (21).

with eighteen different percentage levels: 0%, 1%, 3%, 5%, 7%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, and 75% of Premium 91 gasoline.

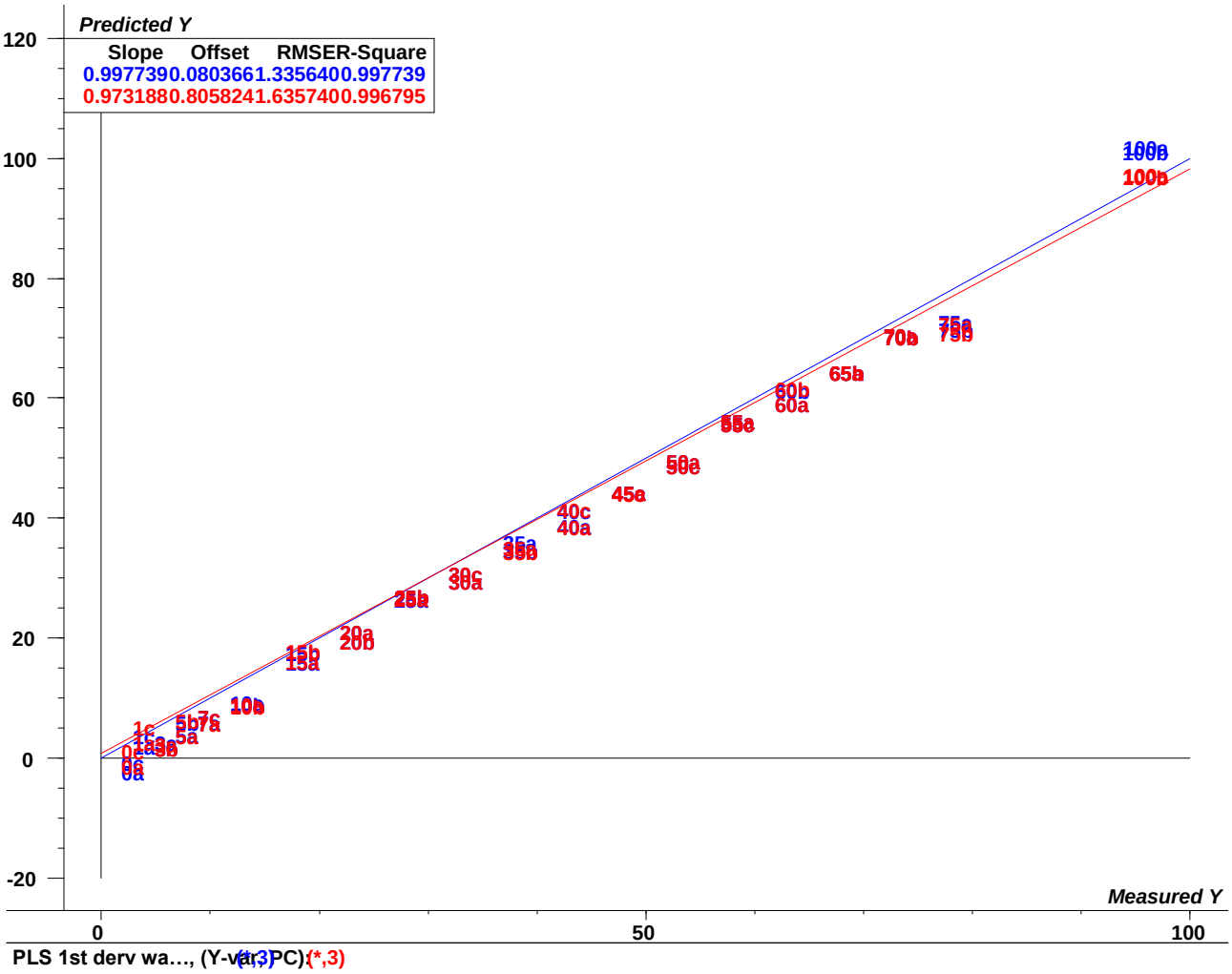


Figure 6. PLS calibration plot for pure super premium 95 and adulterated with premium 91 gasoline samples

It can be seen from Figure 7 that its having small value of RMSECV = 1.33% for 3 factors with R^2 = 99% and of 0.99 correlationship. RMSECV is calculated using Eq. 1:

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

where y_i is the measured value (actual % of adulteration), $\hat{y}_i$ is the % of adulteration predicted by the model, and n is the number of segments left-out in the cross-validation procedure, which

is equal to the number of samples of the training set. Smaller values of RMSECV are indicative of a better prediction ability of the model (20).

Similary the factor loading plot that is analogous to correlation coefficients, by squaring them give the amount of explained variation for PLS model is shown in Figure 7.

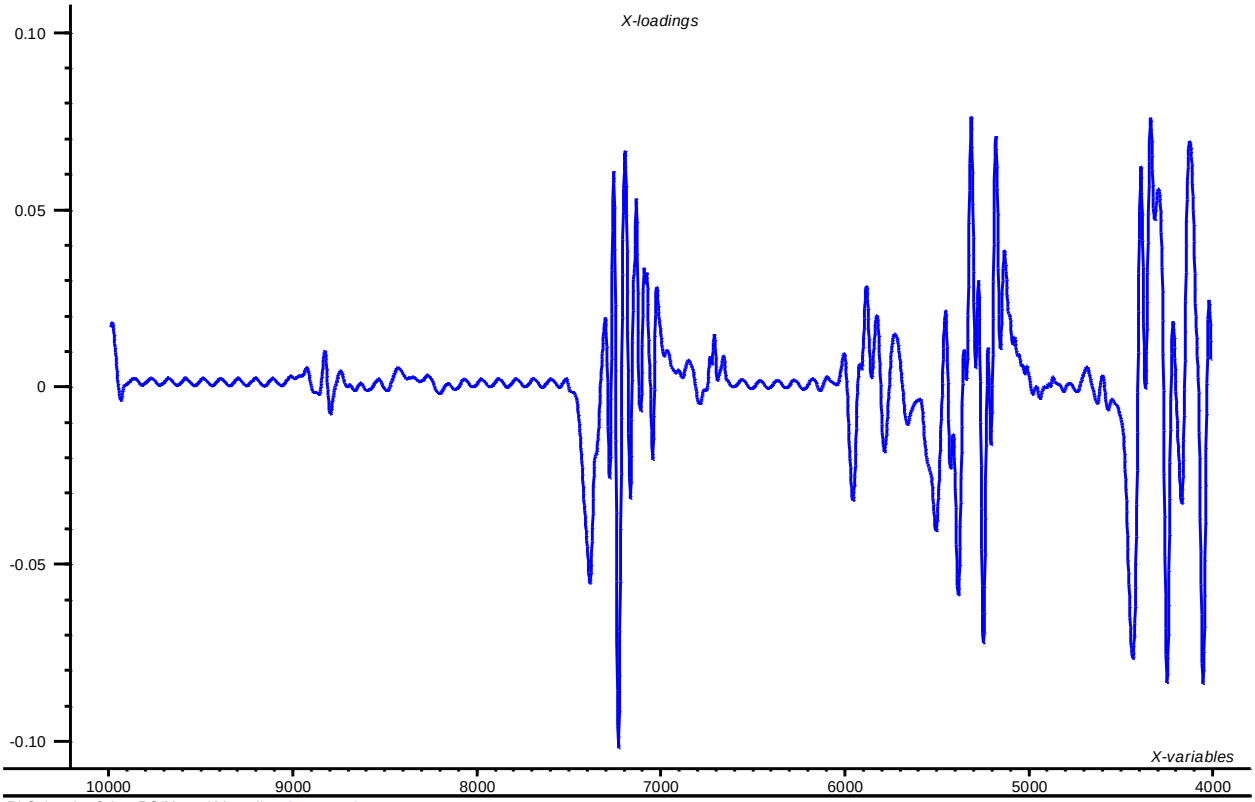


Figure 7. Factor loading plot for factor 1

Figure 7 shows the factor loading plot for factor 1. It tell us how much of the variation in a variable is explained by the factor. In this case, 3 factors contain 77% of the total variation. Factor 1 explains 43 % of the variation, factor 2 explains 26%, and factor 3 explains 8%. The remaining 3 components explain only 21% the remaining two loading plots are not shown here. A residual validation variance plot that shows the number of factors important is shown in Figure 8.

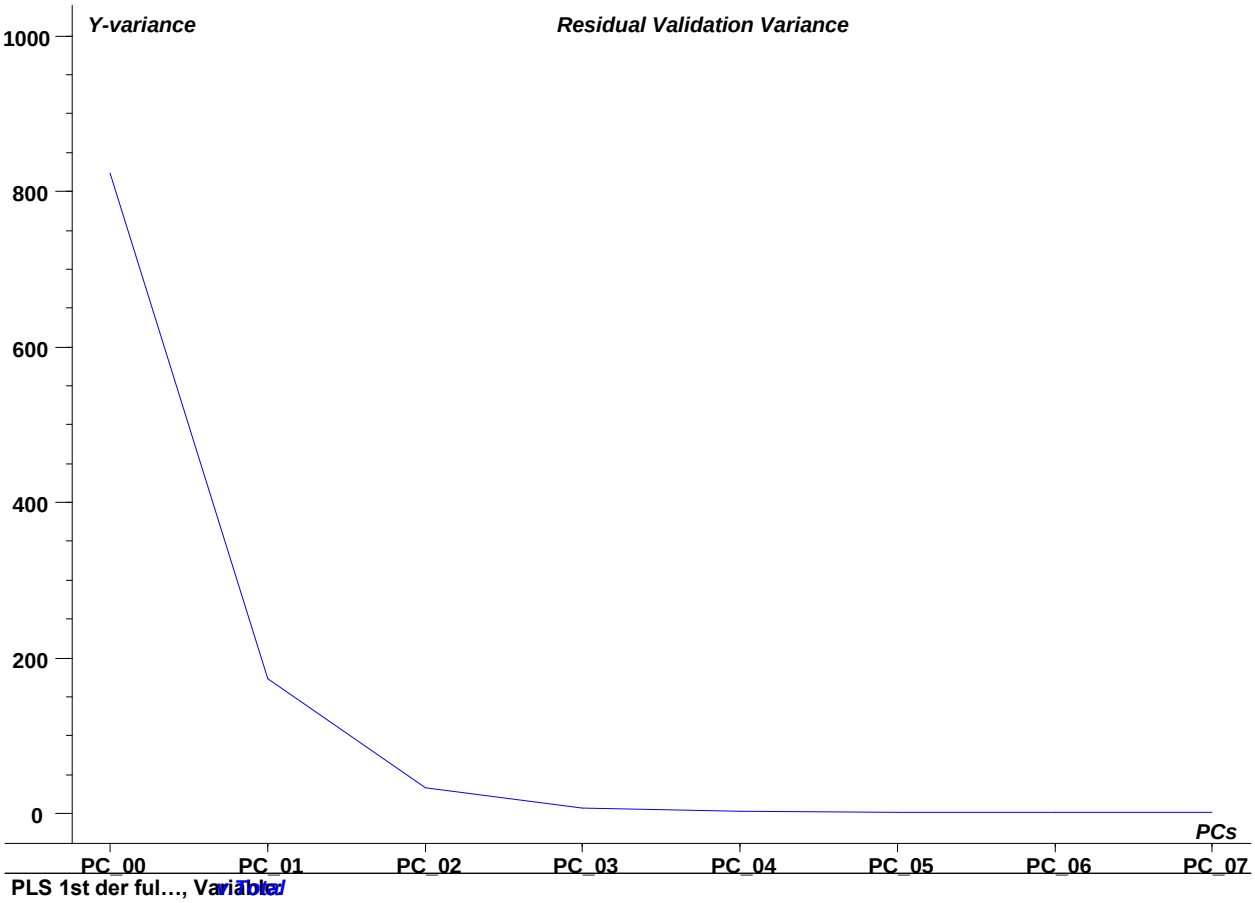


Figure 8. The Residual Validation Variance plot for the PLS regression model

It shows that three components have mostly explained the spectral data.

PLS calibration model was then applied on the test set of the remaining 30% samples to check it prediction ability (described in the experimental section) and its performance is shown in Figure 9.

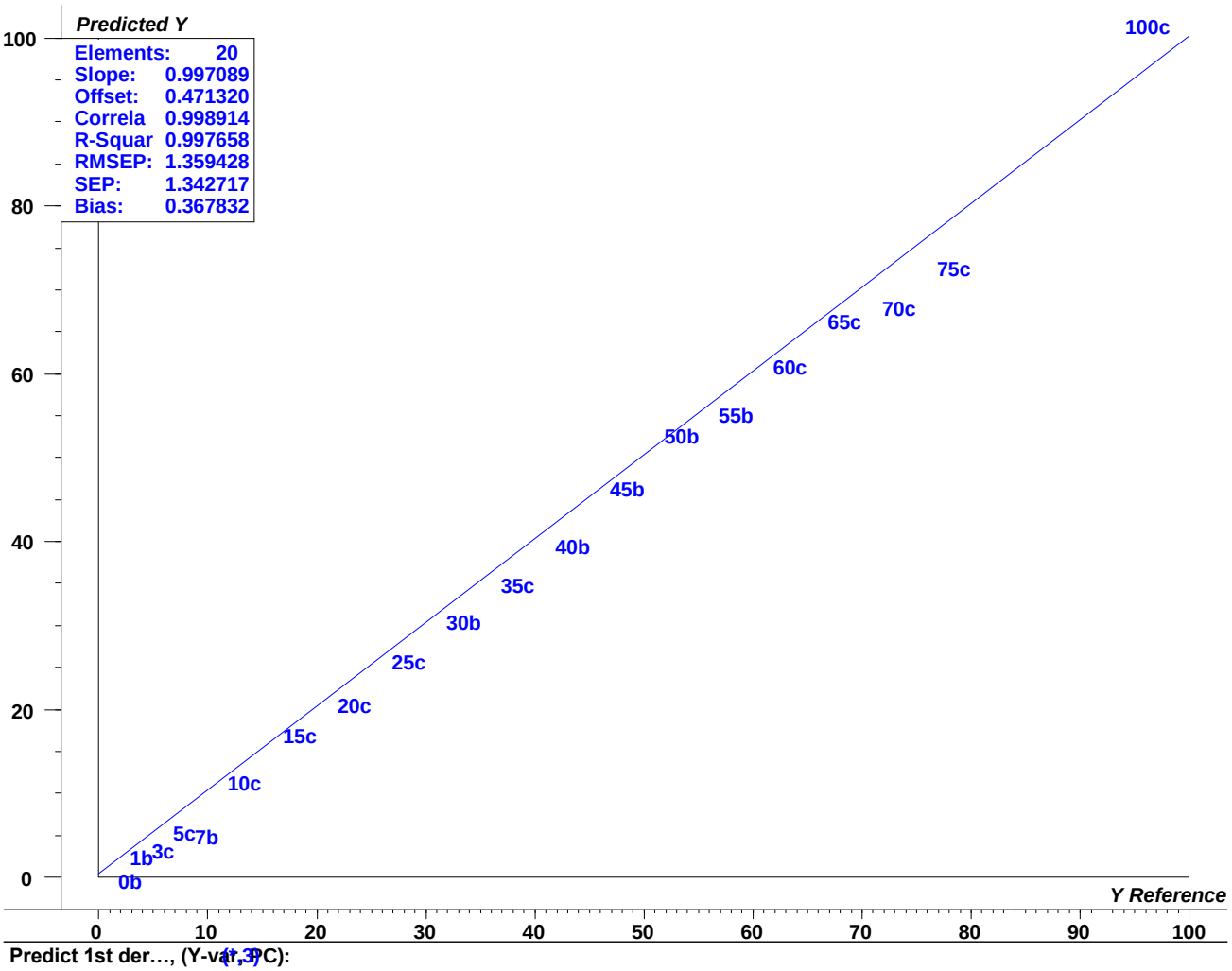


Figure 9. Prediction plot for the 30 % test samples as an external validation set

It can be seen from Figure 9 that the PLS regression model is having a very good prediction abilty with RMSEP value = 1.35% and those 30 % test samples were not used in building the PLS caliberation model. The RMSEP is a statistical measure how well the model predicts new samples (not used when building the model). It is calculated using Eq. 2:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_t}(y_{t,i} - \hat{y}_{t,i})^2}{n_t}} \quad \text{----------(2)}$$

where $y_{t,i}$ is the measured value (actual % of adulteration), $\hat{y}_{t,i}$ is the % of adulteration predicted by the model, and n_t is the number of samples in the test set. RMSEP expresses the average error to be expected in future predictions when the calibration model is applied to unknown samples.

4. Conclusions

It is concluded that this new NIR spectrscopy combined with PCA, PLS-DA and PLS regression models is a suitable technique for detection and quantification of super-premium 95 octane gasoline adulteration with premium 91 gasoline. PLS-DA model can be used as an identifaction tool while PLS calibration models can be used as a quantification tool and it was found that this PLS model is having very good prediction ability and can quantify the lowest level of premium 91 gasoline adulteration less than 1.5 % that is otherwise very difficult to fiind with other convention methods.

References

1. World Bank report – South Asia Urban Air Quality Management. Briefing Note no. 7. July 2002 available at http://www.worldbank.org/saurbanair.
2. World Bank report – Abuses in Fuel markets. Viewpoint Note no. 237. September 2001. Available at http://www.worldbank.org/html/fpd/notes/237/ 237kojim-831.pdf.
3. ASTM – American Society for Testing and Materials. Standard test method for research octane number of spark-ignition engine fuel. West Conshohocken, (PA): ASTM D2699; 2009.
4. ASTM – American Society for Testing and Materials. Standard test method for motor octane number of spark-ignition engine fuel. West Conshohocken, (PA): ASTM D2700; 2009.
5. Moreira LS, d'Avila LA, Azevedo DA. Automotive gasoline quality analysis by gas chromatography: study of adulteration. Chromatographia 2003; 58: 501–505.
6. Wiedemann LSM, d'Avila LA, Azevedo DA. Adulteration detection of Brazilian gasoline samples by statistical analysis. Fuel 2005; 84: 467–473.
7. Takeshita EV, Rezende RVP, Souza SMAGU, Souza AAU. Influence of solvent addition on the physicochemical properties of Brazilian gasoline. Fuel 2008; 87: 2168–2177.
8. Barbeira PJS. Using statistical tools to detect gasoline adulteration. Engenharia Térmica 2002; 1: 48–50.

9. Teixeira LSG, Oliveira FS, Santos HC, Cordeiro PWL, Almeida SQ. Multivariate calibration in Fourier transform infrared spectrometry as a tool to detect adulterations in Brazilian gasoline. Fuel 2008; 87: 346–352.

10. Flumignan DL, Boralle N, Oliveira JE. Screening Brazilian commercial gasoline quality by hydrogen nuclear magnetic resonance spectroscopic fingerprintings and pattern – recognition multivariate chemometric analysis. Talanta 2010; 82: 99–105.

11. Al-Ghoutia MA, Al-Degsb YS, Amera M. Determination of motor gasoline adulteration using FTIR spectroscopy and multivariate calibration. Talanta 2008; 76: 1105–1112.

12. Balabin RM, Safieva RZ, Lomakina EI. Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction Chemometr Intell Lab Syst 2007; 88: 183–188.

13. Balabin RM, Safieva RZ, Lomakina EI. Wavelet neural network (WNN) approach for calibration model building based on gasoline near infrared (NIR) spectra Chemometr Intell Lab Syst 2008; 93: 58–62.

14. Balabin RM, Safieva RZ. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. Fuel 2008; 87: 1096–1101.

15. Balabin RM, Safieva RZ, Lomakina EI. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques Anal Chim Acta 2010; 671: 27–35.

16. Monteiro MR, Ambrozin ARP, Lião LM, Boffo EF, Tavares LA, Ferreira MMC, Ferreira AG. Study of Brazilian gasoline quality using hydrogen nuclear magnetic resonance (^{1}H NMR) spectroscopy and chemometrics. Energy Fuel 2009; 23: 272–279.

17. Balabin RM, Safieva RZ. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. Fuel 2008; 87: 1095–1101.

18. Balabin RM, Safieva RZ, Lomakina EI. Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques. Anal Chim Acta 2010; 671: 27–35.

19. Alexopoulos EC. Introduction to Multivariate Regression Analysis. Hippokratia 2010; 14: 23–28.
20. Fazal Mabood, Farah Jabeen, Manzor Ahmed, Saaida A. A. Al Mashaykhi, Zainb M. A. Al Rubaiey etal "Development of new NIR-spectroscopy method combined with multivariate analysis for detection of adulteration in camel milk with goat milk" Food Chemistry, 221 (2017) 746–750.
21. Svante Wold, Michael Sjostrom, Lennart Eriksson, "PLS-regression: a basic tool of chemometrics" Chemometrics and Intelligent Laboratory Systems, 58 (2001) 109–130.

List of abbreviation

22. • Near Infrared spectroscopy (NIR)
23. • Principle component analysis (PCA)
24. • Partial least discriminant analysis (PLS-DA)
25. • Partial least regression analysis (PLS)
26. • Root mean square error (RMSE)
27. • Root mean square error of cross validation (RMSECV)
28. • Root mean square error of prediction (RMSEP)
29. • Standard Normal Variate (SNV)

Supplementary Material

Click here to download Supplementary Material: Fuel paper figures for revision 2.docx