PCA AND PLS-DA TECHNIQUES IN CONJUNCTION WITH AN FTIR SPECTROMETER FOR ASSESSING THE QUALITY OF GASOLINE FUEL

Lalbiaktluanga[1] And H.H. Thanga[2*]

Abstract
The primary aim of the study is to assess gasoline quality in Mizoram and compare it with samples from neighboring states (Assam) and laboratory kerosene-blended samples. A total of 174 samples were collected, including those from Mizoram and Assam, and laboratory-blended paraffin samples were categorized according to their quality characteristics. Principal Component Analysis (PCA) was used to analyses the data and identify patterns or trends among the samples. However, it seems that PCA alone may not be sufficient for precise classification. The Partial Least SquaresDiscriminant Analysis (PLS-DA) was also employed as a method for classification. PLS-DA is a supervised method that may offer more accurate classification results compared to PCA. The study highlights the importance of ensuring gasoline quality in Mizoram, especially in the face of adulteration practices. The use of advanced analytical techniques like PLS-DA enhances the ability to detect and classify adulterated samples accurately. This research could potentially inform regulatory measures to combat fuel adulteration and ensure consumer safety and satisfaction.

Keywords: Mizoram, Assam, Kerosene, PCA, PLS-DA.

[1,2*]Department of Physics, Mizoram University, Aizawl-796004, Mizoram, INDIA

***Corresponding Author:** H.H. Thanga
*Department of Physics, Mizoram University, Aizawl-796004, Mizoram, INDIA (hthanga@rediffmail.com)

DOI: 10.53555/ecb/2024.13.03.09

Introduction

The Petroleum fuel is a chemical mixture of aliphatic hydrocarbons (linear, chained, cyclo, and polycyclic) and aromatic chemicals. Different chemical compounds have been added to the gasoline in attempt to improve its quality [1]. Petroleum commercial products, such as gasoline fuel, are the primary sources of energy for transportation in Mizoram (northeast India). As a result of increased demand and state government-enforced taxation, fuel adulteration has already become pretty common practice in the localities. It has been suggested that fuel adulteration during transportation, which is sometimes regarded as one of the most common adulteration practices, occurs when it is imported by road into Mizoram from the nearby state. Given the importance of energy in transportation, it is critical to monitor the quality and characteristics of gasoline fuel imported within the city of Aizawl. To prevent fuel adulteration and ensure quality control, the government of India has implemented various measures and actions in accordance with the

Marketing Discipline Guidelines (MDG) and a Dealership Agreement substituting between Retail Outlet (RO) dealers and public sector Oil Marketing Companies (OMCs). OMCs has taken significant steps to improve fuel quality monitoring and surveillance at Retail Outlets. The government received 4145 complaints during the first three years of 2019 from all other states except Mizoram, and 32 retail establishments were shut down nationwide. [2]. Due to the fact that Mizoram has so far not made any initiative steps in comparison to other states, it is essential to classify gasoline samples in order to maintain fuel quality across the state.

Numerous research regarding the usages of spectroscopic methods in combination with multivariate data analyses have been performed. Mabood, F, et al., investigated standard samples of Premium 91 and Super Premium 95 grade gasoline, with the Super Premium 95 samples contaminated with eighteen different percentage levels of 91 premium gasoline [3]. For statistical analysis, NIR spectroscopy with multivariate methods such as PCA, PLS-DA, and PLSR were used and it was concluded that this technique is suitable for detection and quantification of adulterated gasoline samples [3]. Doble et al. did the Gas Chromatography/Mass Spectrometry, Principal Component Analysis and Artificial Neural Network for the classification of premium and regular gasoline from Canadian Petroleum Product Institute [4]. In the spectroscopic method, GC-MS is considered one of the finest and most dependable approaches for analysis. Due to the expensive and time-consuming nature of GC-MS, the straightforward, quick, and affordable technique of FTIR is most frequently employed. As a result, by processing various data, FTIR spectroscopy with Multivariate Analyses can be used to classify the quality of gasoline fuels.

The fuel's quality parameters are essential even in terms of the manufacturing chain, production and marketing in order to certify its quality. Several quality standards necessitated the use of complex technology, qualified employees, and time. This difficulty, however, can be overcome using a Mid-Infrared spectroscopic approach, which can play an essential role in the speedy and effective regulation of petroleum products [5-8]. The commonly used spectroscopic method, Fourier Transform Infrared Spectroscopy (FTIR), is a non-destructive analytical technique that allows for quick and direct sample analysis in order to efficiently manage certain of the characteristics of fuel [9-13]. Based on this, integrating chemometric techniques with infrared spectroscopy has produced promising findings for extracting hidden information from the spectra.

The infrared spectra were categorized using multivariate approaches such as PCA and PLSDA. Among these, PCA techniques aim to pack a lot of data into a compact space; It is a multivariate qualitative projecting method that is unsupervised and relies on a linear array of variables. As a result, PCs are plotted orthogonally to one another to provide the largest spatial separation of the samples in graphical projection [14]. On the other hand, PCA was used to represent the similarities and differences between the samples using a score plot, and the loading plot was used to analyze the results on score plots [15-17]. PLS-DA, on the other hand, is a supervised pattern technique that requires a training set of known groups [18]. It is a calibration algorithm based on the PLS technique, except instead of concentration information, it employs class labels as the dependent y vector [19]. Usually, it is also used as a classification purpose. The

PLS2 algorithm is typically used when there are more than two classes involved in the application. The supervised PLS-DA is found to be more robust than the unsupervised PCA for classification purposes. Thus, the spectroscopic method of FTIR in combination with multivariate data analyses of Principal Component Analysis (PCA) and Partial Least Square-Discriminant Analysis (PLSDA) appears to be the most suitable for Spectral Analysis in the present work. In terms of fuel quality control, the objective of this study is to

compare gasoline fuel obtained in Aizawl (Mizoram), nearby states, and adulterated fuel. Furthermore, the chemical compounds present in the samples are then used to classify the fuel quality.

2. Material and Methods

2.1 Sample collection and Preparation

Indian Oil Corporation Limited (IOC) provided a total of 174 samples: 143 from Aizawl, Mizoram (Northeast India) and 12 from Assam (Northeast India). For comparison, another 19 samples of Laboratory blended Kerosene were used. It is then divided into three categories: Aizawl samples, Assam samples, and Kerosene Blended samples. Due to the unknown composition of all samples, it is used for classification without any purification.

2.2 FTIR Spectral Data Acquisition

The Mid-IR spectra were recorded using an ABB Bomem Spectrometer with Horizon MB3000, an attenuated total reflectance (ATR) sample holder, and a ZnSe crystal. The spectra of both test and model samples were recorded in the 650-3200 cm^{-1} range. The acquisition system has a resolution of 4cm^{-1} and a total of 30 scans.

2.3 Chemometric Procedure

Unscrambler X of CAMO software, version 10.4.1, was used to carry out the Multivariate procedure, as well as various data pre-processing techniques such as baseline correction and standard normal variate (SNV). Baseline adjustments are used to rectify the spectral offset by either adjusting the data to the minimum point in the data or by applying a linear correction based on two user-defined variables. They are mostly employed in spectroscopy. The two changes can be done one at a time or simultaneously. The performance of these techniques was used to modify the spectral offset by adjusting the data to the least point in the data. SNV is used to remove scattered effects by centering and scaling each individual spectrum, and the performance of these techniques was used to adjust the spectral offset by adjusting the data to the minimum point in the data [20].

The unsupervised method of Principal Component Analysis (PCA) was employed to identify the possible outlier, which was chosen for the first category because it is one of the most powerful methods for revealing hidden data structures within the sets of data. It portrays the connection between samples and variables based on their similarities or differences. PCA is also known as the projection method because it projects the information carried by the original variables onto a smaller number of latent variables known as Principal Components (PC). Each PC explains a portion of the entire information included in the original data, and the first PC contains the most information in the data set; the number of PCs can be determined based on the total variances explained.

To produce a better and unambiguous classification, the supervised techniques of Partial Least Square Discriminant Analysis (PLS-DA) is used; in most cases, the supervised approach is used to make predictions. PLS-DA is a well-established and effective method for predicting the number of samples to a predefined group based on a set of measured data. It is also used in classification purposes, where PLS is used to model the differences between several classes. PLSDA, on the other hand, can be defined as a feature of PLS Regression with the aim of predicting a response matrix, Y, from a predictive matrix, X [15]. The Partial Least Squares Regression (PLSR) models the X- and Y-matrices simultaneously in order to find the latent (or hidden) variables in X that best predict the latent variables in Y. These PLS components are comparable to principal components, but they are referred to as factors. In the case of PLS-DA, the PLS model uses a single response variable to code for class label, as shown below: Members of one class are assigned a 1, while members of the other are assigned a +1. If there are three or more classes, the model assigns one response variable to every class (-1/+1 or 0/1, which is equivalent). The model then includes a number of Y-variables. To evaluate the models' performance, four parameters were calculated: the root mean square error (RMSE), R-Squared, Slope, and Bias. The RMSE and RSquared values characterise the precision and accuracy of the predictions, and they indicate how well future predictions for a given set of variables could be predicted. Bias values that are significantly greater than zero indicate systematic errors, and the closer the slope is to one, the better the data can be modelled [21, 22].

3. Result and Discussion

3.1 Measurement of Spectral Properties

Gasoline fuel is composed of various types of hydrocarbons that contain aromatic compounds. In general, the presence of chemical components in gasoline compounds has a direct influence on spectrum characteristics. As a result, variations in chemical composition can have an impact on the refining process, fuel quality parameters, marketing, and environmental pollution. As a

consequence, the characterization method is important in the petroleum industry.

Figure 1 shows a comparison of the ATR-FTIR spectra of gasoline and kerosene blended samples. The bands observed between 3200 and 2700 cm^{-1} correspond to alkyl C-H stretching vibrational modes, which are followed by in-plane C-H bending and C-C modes between 1600 and 1000 cm^{-1}, and the bands below 1000 cm^{-1} correspond to aromatic compounds' C-H out-ofplane (oop) modes [23].

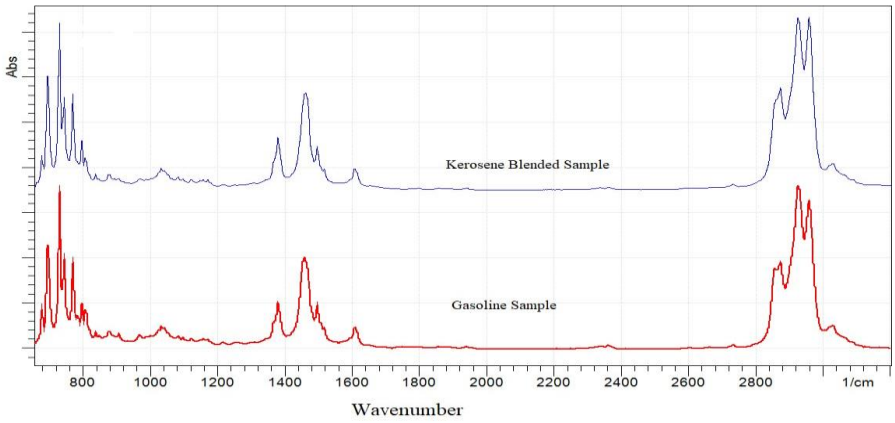


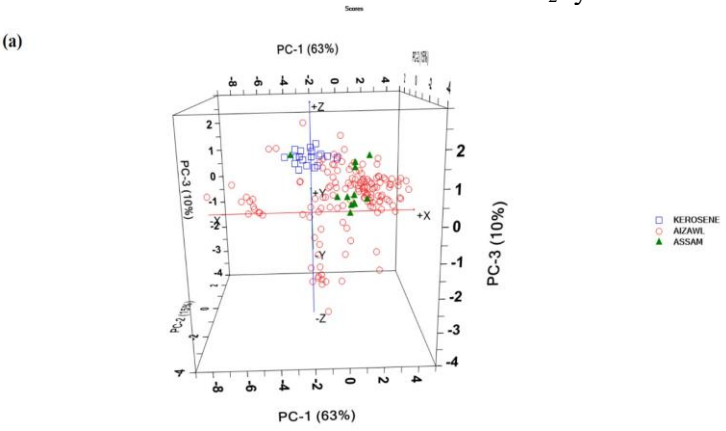
Figure 1: Full Spectra of Gasoline and Kerosene blended samples at the mid IR region of 650-3200 cm^{-1}

3.2 Processing of Chemometric Data

Chemometric analysis is performed on the mid-IR spectra of all samples ranging from 650 to 3200 cm^{-1}. To observe possible separation and comparison, analysis was performed separately at the spectral region of 2800-3200 cm^{-1}, where Kerosene blended sample shows major peaks at this region.

3.2.1 Principal Component Analysis:

Figure 2 illustrates the PCA 3D score and correlation loading plot of the 650-3200 cm^{-1} region, i.e., the entire region. In the score plot of Figure 2(a) the first three Principal Component explained 88% of the total variability i.e., PC1 63%, PC2 15% and PC3 10%. Examining the score plot in Fig. 2 in greater detail, it is difficult to distinguish the gasoline sample due to the unclear variation in the Principal Components. Along the positive side of PC1, we discovered that a group of kerosene blended samples is in the same group as a large number of Aizawl samples. The variation of the kerosene blended sample and large number of Aizawl samples, according to the correlation loading plot of Figure 2(b), it is primarily due to the stretching vibration of aromatic ring modes at 1425, 1504 cm^{-1} and the vibration of aromatic compounds at 802, 835, and 873 cm^{1}. At 1002, 1058, and 116 cm^{-1}, there is also some contribution from C-H bond bending vibration in the plane of the benzene ring. Aside from that, a small number of Aizawl samples are found on the negative side of PC1, owing to vibration of aromatic compounds at 690 and 730 cm^{-1}. Along with this, the aromatic C-H bending and stretching vibrations at 1031, 1494, 1604 and 3028 cm^{-1} contribute significantly to this variation. A few Assam as well as Aizawl samples are found to be on the negative side of PC2, which is primarily caused by C-CH_2 scissor bending vibration at 1488
cm^{-1} and CH_2 symmetric vibration at 2852 cm^{-1}.



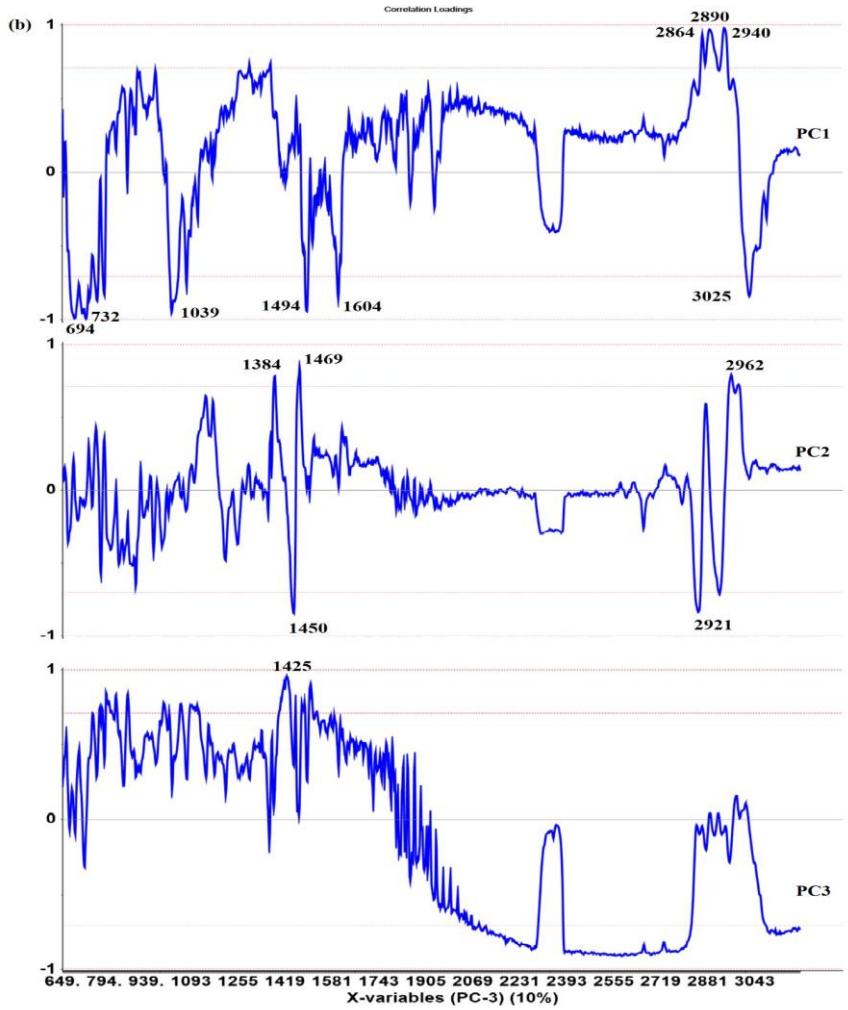


Figure 2: (a) 3D-Score plot and (b) Correlation Loading Plot of PCA

The separation of Aizawl and Assam samples could not be found using the above Principal Component Analysis, and the kerosene blended samples were always found within the major group of samples. While PCA can help in visualizing and understanding the structure of the data, it may not always lead to optimal class separation, especially if the classes are not well-separated in the original feature space. In such cases, more sophisticated supervised classification techniques might be necessary to achieve better classification performance. However, PCA can be a useful preprocessing step in classification tasks, it is not a classification technique itself. Supervised classification methods are typically needed for accurate class differentiation, especially in cases where class separation is not clear in the original feature space.

3.2.2 Partial Least Square Discriminant Analysis (PLS-DA)

The PLS-DA method, like the PCA classification, depicts Factor rather than Principal Component. Figure 3 shows the PLS-DA Score and correlation Loading plot. As from score plot in Figure 3(a), we could see that the Aizawl samples are mainly kept separate towards the positive side of Factor-1, which is primarily due to aromatic stretching vibrations at 686, 736, 771, and 794 cm^{-1} as shown in the loading plot of Figure 3(b). Additionally, the aromatic scissor bending vibrations at 1031, 1494 cm^{-1}, as well as the aromatic C-H bending vibration at 3028 cm-1, contribute to the variation towards Factor-1. The kerosene blended samples are separated mostly on positive side of Factor-2, due mainly to methylene structure unit of C-CH_2 scissor bending vibrations at 1400 and 1475 cm^{-1}, that has a longer carbon chain. We also noticed that a few samples from Aizawl and one from Assam were detached along with kerosene blended samples.

For kerosene-blend samples, PLS-DA provides a clear variation. However apart from this 196 variability, the major part of the samples are associated with a same large group, and no other 197 possible variation on this region can be found.

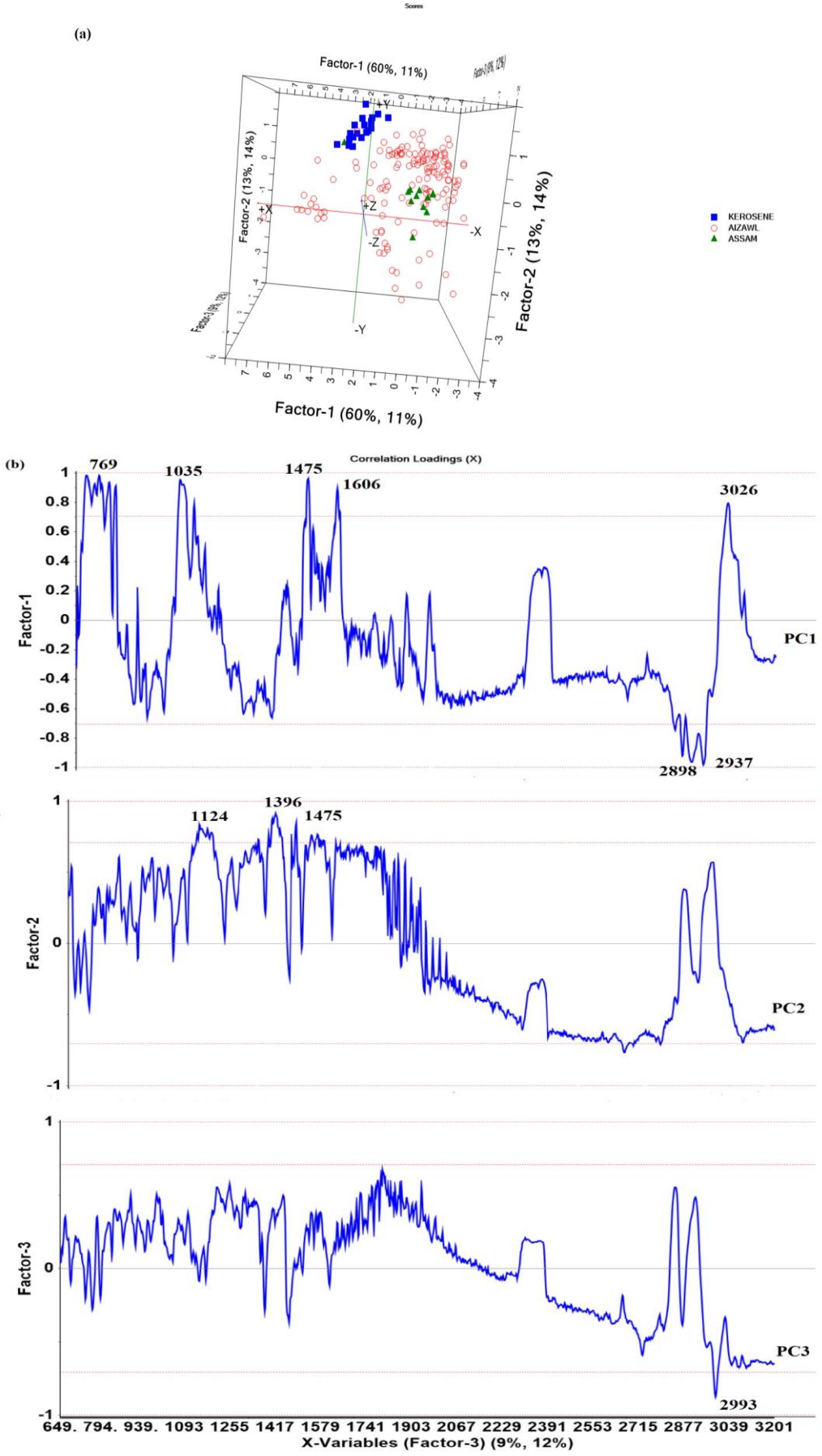


Figure 3: (a) 3-D Score plot and Loading Plot of PLS-DA classification analysis

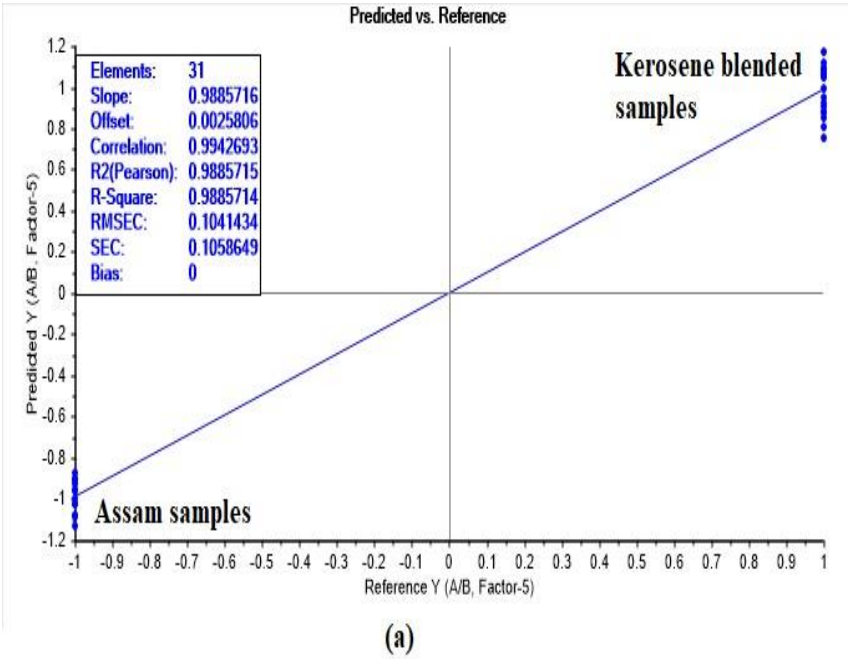
In the context of kerosene blended samples, where the goal is likely to differentiate between different blends or compositions, PLS-DA can be particularly effective in identifying spectral or chemical features that are discriminative for each blend. It can help in quality control, authentication, and monitoring of kerosene blends in various industries, including aviation, automotive, and energy sectors. Overall, PLS-DA can provide clear variation and detailed information for classification analysis, especially in complex datasets with interrelated variables such as those found in chemical analysis.

3.2.3 PLS-DA Prediction Analysis:

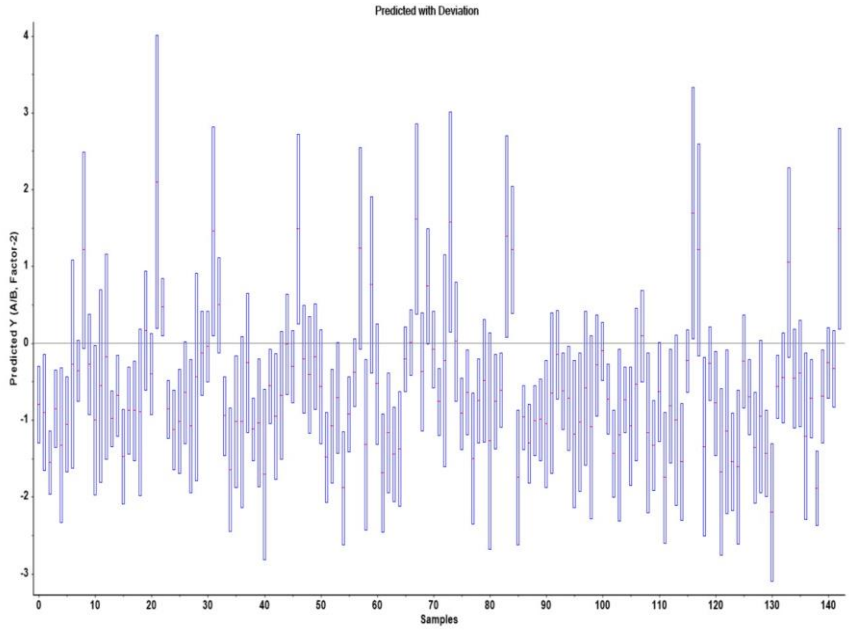
It seems that the PLS-DA modelling in this study effectively classified samples into three classes: Kerosene Blended, Assam samples, and samples that were neither blended with kerosene nor from Assam. The PLS-DA model considered Kerosene Blended and Assam samples as model samples, while Aizawl samples served as test samples for validation. The performance of the model demonstrated high accuracy, with values such as RMSEC of 0.1041 and R-Square of 0.9886, indicating the model's reliability in predicting class membership as shown in Figure 4(a).

According to the prediction result from Figure 4(b), the samples those falling within the projected values of 1 and 2, considered as laboratory-blended gasoline samples (model A), and those falling between -1 and -3, classified as Assam samples (model B). Samples near the prediction value of 0 were not classified into either of these two classes. So that, in accordance with PLS-DA prediction results, 35 additional test samples were classified as belonging to class B (Assam samples), while 14 additional test samples were classified as belonging to class A (laboratory-blended gasoline samples). Test samples near the value of 0 were classified as neither class A nor class B.

The PLS-DA prediction model effectively split test samples into three classes: Kerosene Blended, Assam samples, and samples that did not belong to either of these two categories. In summary, the PLS-DA prediction model demonstrated robust performance in classifying test samples into distinct categories based on their chemical composition and origin. This classification approach provides valuable insights into the quality and composition of fuel samples, aiding in quality control and regulatory efforts within the region of study.



(a)



(b)

Figure 4: (a) The predicted vs Reference plot and (b) Predicted result of PLS-DA analysis**.**

From the above two analysis, PCA does not clearly separating the Aizawl samples and comparative samples from nearby states with kerosene blended samples. The presence of aromatic chemicals in the fuel samples may indeed lead to two significant groups of variation. However, it seems that these groups are not clearly corresponding to the geographical origin or the presence of kerosene components in the samples. This could be due to several factors such as variations in the blending process, source of the fuel, or other chemical additives present in the samples. It appears that distinguishing between Aizawl samples containing kerosene components and those from other nearby states is challenging based solely on the PCA analysis and the presence of aromatic chemicals. This highlights the complexity of the samples and the need for more sophisticated analytical techniques or additional information to differentiate between them effectively. In such cases, alternative methods like Partial Least Squares Discriminant Analysis (PLS-DA) or other supervised classification techniques may be more suitable for identifying the key features that differentiate between the samples. PLS-DA, in particular, can leverage the class labels (e.g., geographical origin or kerosene content) to build a more discriminative model.

However, The PLS-DA analysis effectively segregates kerosene mixed samples into a distinct group due to their longer carbon chains compared to samples acquired from filling stations. This indicates that PLS-DA is sensitive to the differences in chemical composition and can identify samples with longer carbon chains, characteristic of kerosene. A small number of Aizawl samples are found to have longer carbon chains, suggesting potential tampering or adulteration with kerosene. Additionally, the presence of unidentified adulterants in some test samples further supports the notion of tampering or poor sample quality in the Aizawl area. PLS-DA classification reveals that Aizawl samples can be divided into two classes, with some samples resembling kerosene mixed samples and others resembling Assam samples. This indicates variability in sample quality and composition within the Aizawl area. The presence of unidentified adulterants in some Aizawl samples suggests the possibility of additional chemical substances besides kerosene being used as adulterants. This highlights the importance of thorough analysis and detection methods to identify potential adulterants in fuel samples.

CONCLUSION

The present work mentioned assessing the quality of gasoline samples within the state of Mizoram, northeast India, using mid-infrared spectra and employing PCA and PLS-DA techniques for classification. It addresses the importance of assessing the quality of gasoline samples within Mizoram, highlighting the need for effective quality control measures in the region's fuel supply. The study finds that while the PCA approach does not provide easily recognizable variation among all the samples, the PLS-DA method effectively separates the kerosene blended sample and provides precise information. The PLS-DA approach also achieves high classifier accuracy on

test samples. It suggests that in practice, PLS-DA approaches offer more reliable categorization information than PCA techniques. This implies that PLS-DA is better suited for discriminating between different fuel samples and identifying potential adulteration or quality issues.

Overall, the study highlights the significance of employing advanced analytical techniques like PLS-DA for quality assessment and classification of fuel samples. The findings underscore the importance of adopting robust methods to ensure the integrity and quality of fuel supplies, particularly in regions like Mizoram where maintaining fuel quality is crucial for various sectors and industries.

Acknowledgement

The Science and Engineering Research Board (SERB), Department of Science and Technology (DST), New Delhi, India (EMR/2016/005525), provided funding for this work.

References

1. James, G.S., Feedstock Composition, in Handbook of Petroleum Refining. 2016, CRC
a. Press.
2. https://pib.gov.in/newsite/PrintRelease.aspx?relid=195063
3. Mabood, F., et al., Detection and estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with multivariate methods. Fuel, 2017. 197: p. 388-396.
4. Doble, P., et al., Classification of premium and regular gasoline by gas chromatography/mass spectrometry, principal component analysis and artificial neural networks. Forensic Sci Int, 2003. 132(1): p. 26-39.
5. Barra, I., et al., FTIR fingerprints associated to a PLS-DA model for rapid detection of smuggled non-compliant diesel marketed in Morocco. Vibrational Spectroscopy, 2019. 101: p. 40-45.
6. Moros, J., S. Garrigues, and M.d.l. Guardia, Vibrational spectroscopy provides a green tool for multi-component analysis. TrAC Trends in Analytical Chemistry, 2010. 29(7): p. 578-591.
7. Pasquini, C., Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications
a. Chem. Soc, Apr 2003: p. 14 (2).
8. Inan, T.Y., A. Al-Hajji, and O.R. Koseoglu, Chemometrics-Based Analytical Method Using FTIR Spectroscopic Data To Predict Diesel and Diesel/Diesel Blend Properties. Energy & Fuels, 2016. 30(7): p. 5525-5536.
9. Fernandes, D.D.S., et al., Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection. Talanta, 2011. 87: p. 30-34.
10. Oliveira, F.C., et al., Adulteration of diesel/biodiesel blends by vegetable oil as determined by Fourier transform (FT) near infrared spectrometry and FT-Raman spectroscopy. Anal Chim Acta, 2007. 587(2): p. 194-9.
11. Guimarães, E., et al., Infrared Spectroscopy and Multivariate Calibration for Quantification of Soybean Oil as Adulterant in Biodiesel Fuels. Journal of the American Oil Chemists' Society, 2015. 92(6): p. 777-782.
12. Gontijo, L.C., et al., Quantification of soybean biodiesels in diesel blends according to ASTM E1655 using mid-infrared spectroscopy and multivariate calibration. Fuel, 2014. 117: p. 1111-1114.
13. Durazzo, A., et al., Qualitative Analysis of Traditional Italian Dishes: FTIR Approach. Sustainability, 2018. 10(11).
14. Santos, V.H.J.M.d., et al., Discriminant Analysis and Cluster Analysis of Biodiesel Fuel Blends Based on Fourier Transform Infrared Spectroscopy (FTIR). Energy & Fuels, 2016. 30(6): p. 4905-4915.
15. Barra, I., et al., Discrimination of diesel fuels marketed in Morocco using FTIR, GC-MS analysis and chemometrics methods. Talanta, 2020. 209: p. 120543.
16. Kharbach, M., et al., Characterization and classification of PGI Moroccan Argan oils based on their FTIR fingerprints and chemical composition. Chemometrics and Intelligent Laboratory Systems, 2017. 162: p. 182-190.
17. Sudol, P.E., et al., Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis. Talanta, 2020. 206: p. 120239.
18. da Silva, M.P.F., et al., Classification of gasoline as with or without dispersant and detergent additives using infrared spectroscopy and multivariate classification. Fuel, 2014. 116: p. 151-157.
19. Silva, A.C., et al., Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods. Talanta, 2012. 93: p. 129-134.
20. Wang, Y., D.J. Veltkamp, and B.R. Kowalski, Multivariate instrument standardization.

Analytical Chemistry, 1991. 63(23): p. 2750-2756.
21. Lavine, B., A user-friendly guide to multivariate calibration and classification, Tomas Naes, Tomas Isakson, Tom Fearn and Tony Davies, NIR Publications, Chichester, 2002, ISBN 0-9528666-2-5, £45.00. Journal of Chemometrics, 2003. 17(10): p. 571-572.
22. da Silva, N.C., et al., Classification of Brazilian and foreign gasolines adulterated with alcohol using infrared spectroscopy. Forensic Sci Int, 2015. 253: p. 33-42.
23. Smith, B., Infrared spectral interpretation: a systematic approach. 2018: CRC press.