Development of methods based on NIR and Raman spectroscopies together with chemometric tools for the qualitative and quantitative analysis of gasoline

Thesis for the European Master's degree in Quality in Analytical Laboratories

by:

Ricardo Henrique de Paula Pedroza

Supervisors:

Dr. Aaron Urbas, *NIST*, United States of America

Prof. Dr. Bjørn Grung, *University of Bergen*, Norway

Prof. Dr. Werickson Fortunato de Carvalho Rocha, *INMETRO*, Brazil

Bergen, September 2019

Erasmus Mundus

emQal

"Success is no accident. It is hard work, perseverance, learning, studying, sacrifice, and most of all, love of what you are doing or learning to do."

Edson Arantes do Nascimento (Pelé).

Acknowledgements

I would like to express heartfelt and gratitude to my supervisors. Dr. Aaron Urbas from NIST, for all effort to obtain gasoline samples for the project, his many advices in how to proceed with the spectroscopic analyses and chemometrics. Prof. Dr. Werickson Rocha from INMETRO for his advices in chemometrics and for suggesting many ideas to apply in this project. Prof. Dr. Bjørn Grung, whom provided me knowledge and learning in chemometrics which were essential to perform the data analysis of this project.

I am completely grateful to the financial support provided by the European commission for providing me this grant and the opportunity of performing this master's degree. I am also thankful to all people involved in the EMQAL organization, specially to prof. Dr. Miquel Esteban who introduced me to this master a few years ago.

Many thanks to the NIST for opening its doors for the execution of this research project in collaboration with EMQAL, and also for the financial support. I would also like to thank to people from the MML group, specially to Katrice and Laurell who welcomed us very well at NIST and have always been very attentive to us.

I would like to thank my parents, Ricardo Augusto and Maria Ivoneide, for all encouraging and motivating me.

I would like to demonstrate my gratitude to my girlfriend. Thanks for being patient, believing in our relationship even with the distance during this period, and for her moral and sentimental support.

Last and no least, I acknowledge to my all my friends and EMQAL colleagues who collaborated in some way for this experience, it would not have been the same without you.

Thank you,

Ricardo H. P. Pedroza

ABSTRACT

Gasoline quality control is essential for SI engines performance and to reduce environmental impacts by generation of undesirable pollutants. Methods established by the American Society for Test and Materials (ASTM) are the most employed for determining physicochemical quality parameters of motor gasoline, however, these methods present some disadvantages such as time-consuming analysis and need of large amount of sample. For this purpose, near-infrared (NIR) and Raman spectroscopies could be promising alternatives, since they are nondestructive techniques which require little or no sample preparation, a small amount of sample, short analysis time, and also present the possibility of simultaneous determination of many parameters. Although, the use of chemometric tools is often needed in order to extract maximum of useful information from the NIR and Raman spectra related to the parameter being studied. In this work, the qualitative classification of commercial gasoline samples related to their ethanol contents and antiknock indexes was reached by using principal component analysis (PCA) and soft independent modelling of class analogy (SIMCA) models. The values for the misclassification error obtained for the classification of these parameters by both NIR and Raman spectroscopies were less than 3.0%. The multivariate calibration technique, partial least squares (PLS), was used for both NIR and Raman data to obtain predictive models for the quantification of eight physicochemical quality parameters of gasoline such as relative density, motor octane number, research octane number, antiknock index, and gasoline composition by aromatics, benzene, olefins, and paraffins. The accuracy of these PLS models was evaluated by applying the elliptical joint confidence region (EJCR) test, and the ideal theoretical point (slope=1, intercept=0) was involved by the ellipses of all obtained models by both NIR and Raman data demonstrating that BIAS is absent in a confidence interval of 95%. The results obtained in this study demonstrated that both spectroscopic techniques together with chemometric tools provided an excellent performance, thus, being good alternatives to the conventional methods to be used for the quality control of motor gasoline.

ACRONYMS AND ABREVIATIONS

AKI: Antiknock Index

ARE: Average Relative Error

ASTM: American Society for Tests and Materials

BBI: Broadband Inverse

CAL: Calibration

CFR: Cooperative Fuel Research

CV: Cross-Validation

EJCR: Elliptical Joint Confidence Region

FACE: Fuels for Advanced Combustion Engines

FIA: Fluorescent Indicator Absorption

GC: Gas Chromatography

^{1}H NMR: Proton Nuclear Magnetic Resonance

IR: Infrared Radiation

LV: Latent Variable

MC: Mean Center

ME: Misclassification Error

MON: Motor Octane Number

MSC: Multiplicative Scatter Correction

NAFS: North American Fuel Survey

NIR: Near Infrared

NIRS: Near Infrared Spectroscopy

NMR: Nuclear Magnetic Resonance

PC: Principal Component

PCA: Principal Component Analysis

PLS: Partial Least Squares

PREC: Precision

R: Reflectance

RMSE: Root Mean Square Error

RMSEC: Root Mean Square Error of Calibration

RMSECV: Root Mean Square Error of cross validation

RMSEP: Root Mean Square Error of Prediction

RON: Research Octane Number

RSD: Relative Standard Deviation

SEN: Sensitivity

SI: Spark Ignition

SIMCA: Soft Independent Modelling of Class Analogy

SNV: Standard Normal Variate

SPE: Specificity

SRM: Standard Reference Material

T: Transmittance

TMS: Tetramethylsilane

CONTENTS

1. INTRODUCTION

Gasoline is a transparent volatile flammable liquid specifically formulated for Spark Ignition (SI) engines. It is a complex mixture of hundreds of hydrocarbon molecules represented mainly by the broad classes of paraffins, olefins and aromatics, with chains ranging from 4 to 12 carbon atoms per molecule, resulting in a fuel with a boiling point range of approximately 25-225 °C. In the late 1800s, gasoline was produced from the direct distillation of crude petroleum, and there were no test methods or specifications for the formulation. By the 1920s due to the development of first test methods for gasoline and the growing demand of more powerful and efficient engines, gasoline started to be carefully formulated and processes like cracking, reformulation and isomerization have been used to raise the yield of gasoline from crude petroleum [1–5].

The United States, the world's largest consumer of gasoline, was responsible for the average consumption of 275,704 thousand barrels per month for the period between January and May of 2019, according to data provided by the U.S. Energy Information Administration (EIA) [6]. This level of consumption points to the need for gasoline quality control for SI engines performance and to reduce the environmental impacts from the generation of undesirable pollutants. Most of the methods for determining physicochemical quality parameters of gasoline are established by the American Society for Test and Materials (ASTM). Examples of some of commonly used ASTM methods for gasoline which are relevant for this study include the following: ASTM D-2699 [7] and D-2700 [8] for determining research (RON) and motor (MON) octane numbers, respectively; ASTM D-1319 [9] for the determination of saturates, aromatics and olefins of the gasoline by fluorescent indicator absorption (FIA); and ASTM D-4052 [10] to determine the relative density of gasoline.

However, many of these ASTM methods present significant disadvantages. For instance, the determination of RON and MON requires the use of a Cooperative Fuel Research (CFR) engine, which is an expensive instrument, involves a time-consuming analysis (~20-30 min per sample), uses a large amount of sample (~500 mL per analysis) [11] and requires a well-trained operator. The FIA method for hydrocarbon class quantification has a limited scope, lengthy analysis time and a number of potential source of systematic errors that generate results with a broad bias [12]. In this context, there is a need for methods which can overcome those disadvantages.

Near-infrared (NIR) and Raman spectroscopy methods have been increasingly applied in many different areas such as pharmaceuticals [13–16], food and beverages analysis [17–22], disease diagnosis [23,24], petroleum [25–27], among others [28–30]. This growth in applications is mainly due to the many advantages that these techniques offer such as non-destructive analysis, small sample requirements, minimal or no sample preparation, rapid acquisition of spectra, relatively inexpensive instrumentation, and the possibility of simultaneous determination of multiple properties. Nonetheless, these spectroscopic techniques are not without their own drawbacks. NIR spectral bands are typically very broad and highly overlapped. Raman spectra can be significantly impacted by background fluorescence signals. There are also effects related to the presence of systematic errors which are observed as spectral fluctuations among different samples resulting in the presence of noise, scattering and background. The presence of these undesirable spectral contributions require the use of spectral preprocessing tools to minimize or eliminate unwanted signals or enhance signals of interest followed by the application of chemometric methods to extract the maximum amount of useful information from the spectral data that is related to the parameter/property being studied.

In the literature there is a large number of reported methods related to the use of NIR or Raman spectroscopies together with chemometric tools for both qualitative and quantitative analysis of gasoline. The reported methods include applications for determining the following gasoline properties: ethanol content [5,31–37]; density [38–40]; gasoline composition [38,39,41,42]; methanol content [5,33]; research and motor octane numbers [42–45]; Reid vapor pressure [43,44]; distillation temperatures [39,40,46]; and oxygenate concentration (MTBE, ETBE) [32,47]. These publications demonstrate the potential which NIR and Raman spectroscopies together with chemometric tools can provide as alternatives to the standard methods for different analyses of gasoline.

2. OBJECTIVES

This work aims to develop alternative methods for the qualitative and quantitative analysis of motor gasoline based on its chemical and physical parameters. That is performed by using techniques hereby represented by near-infrared and Raman spectroscopies which require little or no sample preparation, a small amount of sample, and very little analysis time. In order to achieve that goal, chemometric tools were applied in these data according to the following specific objectives:

- Use principal component analysis (PCA) to observe whether anomalous samples are present or not on both NIR and Raman data.
- Develop classification models by using PCA and soft independent modelling of class analogies (SIMCA) in NIR and Raman data in order to discriminate commercial gasoline samples based on their ethanol contents and antiknock indexes.
- Develop quantification models based on the use of partial least squares (PLS) regression applied to NIR and Raman data for quantifying eight gasoline parameters represented by: Research Octane Number, Motor Octane Number, Antiknock index, relative density, general hydrocarbon class composition (aromatics, olefins and paraffins) and benzene.
- Perform a comparison between the qualitative and quantitative results obtained using NIR and Raman spectroscopies.

3. THEORY

3.1. Spectroscopy

3.1.1. Near Infrared Spectroscopy

Infrared radiation was discovered by Sir William Herschel in 1800. Sir William was performing an experiment aiming to identify what color from sunlight was responsible for carrying the heat. The experiment was carried out using a glass prism and a thermometer, and it was realized that none of the visible radiation was carrying the heat. Instead, an invisible radiation found just beyond the red radiation was responsible. This invisible radiation was later named infrared [48].

Later, the infrared spectral region was split into three principal sub-regions according to the main applications. The near-infrared was mostly employed for quantitative analysis, the mid-infrared was largely used for qualitative organic analysis and structures determination, and the far-infrared was mainly used for inorganic studies [49].

Infrared radiation (IR) is not energetic enough in to promote electronic transition, however, some molecules present small energy differences between their vibrational states, thus, being able to absorb IR radiation [48]. The requirement for those molecules to absorb IR radiation is that the vibration results in a net change in their dipole moment. If the radiation frequency matches with the natural vibrational frequency of the molecule, the radiation is absorbed and a vibrational mode is induced in the molecule [49].

Near infrared spectroscopy (NIRS) is a vibrational spectroscopic technique based on the electromagnetic radiation located between the visible and mid-infrared regions of the electromagnetic spectrum with a wavelength range of approximately from 780 nm to 2500 nm [49] as shown in figure 1. Molecular absorption of electromagnetic radiation of in this region of the spectrum consists mainly of overtones and combinations of fundamental vibrational modes in the mid-infrared region (2500-5000 nm) [48], mostly associated with X-H bonds, e.g., N-H, C-H, and O-H.

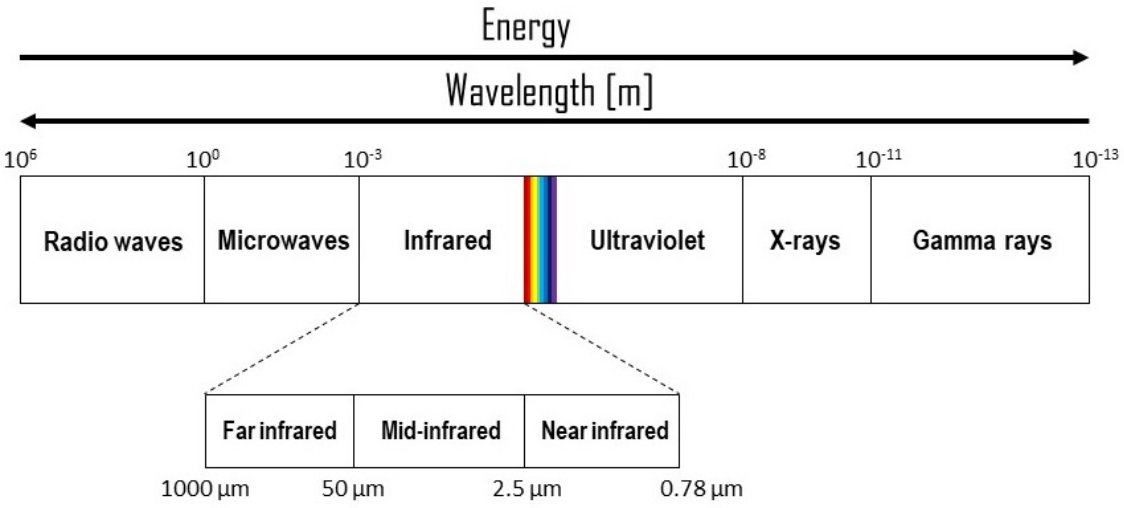

Figure 1. Electromagnetic spectrum with the infrared region highlighted.

The amplitude of the absorption of NIR radiation at any frequency in a molecule is related to its molar absorptivity and the number of molecules found in the beam path of the measuring instrument. This relationship is described by the Beer's law: the absorbance of an analyte is proportional to the molar absorptivity (ε), the path length (b), and the analyte concentration (c) as demonstrated in equation 3.1 [50,51].

$$A = \varepsilon \cdot b \cdot c \qquad \textbf{(3.1)}$$

There are three primary measurement modes of NIR spectroscopy as shown in figure 2. Transmission is the most conventional mode, where a beam of intensity I_0 passes through a sample, where absorption by one or more analytes occurs, and arrives to the detector with intensity *I*. From the ratio of these two measurements the transmittance, T (I/I_0), is calculated. Diffuse reflectance is based on the reflection, R (I/I_r), and the transflection mode is related to a combination of transmittance and reflectance of the radiation. Equation 3.2 demonstrates how the absorbance is calculated for the different modes of NIR spectroscopy. Transmission and transflection are most employed for liquid and semisolid samples, while, diffuse reflectance is most used for solid samples analysis [51].

$$A = \log_{10}\left(\frac{1}{T}\right) = \log_{10}\left(\frac{I_0}{I}\right) = \log_{10}\left(\frac{1}{R}\right) = \log_{10}\left(\frac{I_R}{I}\right) \qquad \textbf{(3.2)}$$

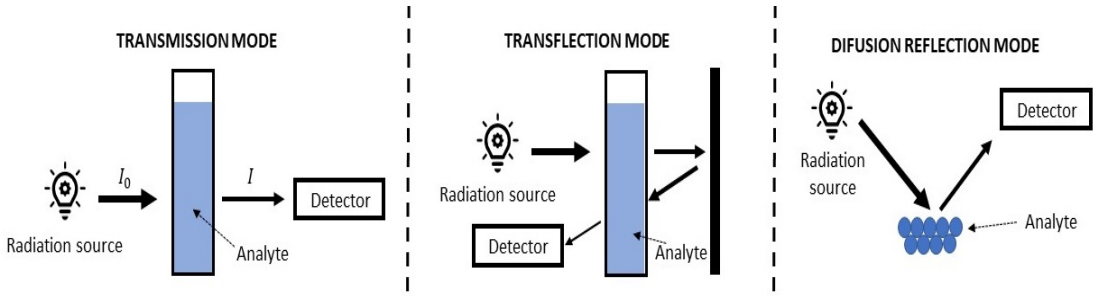

Figure 2. Modes of performing NIR spectroscopy.

3.1.2. Raman spectroscopy

Raman scattering is also a form of molecular spectroscopy which is based on vibrational transitions that can be used to obtain information about the structure and properties of molecules. Raman is most indicated for symmetric vibrations of non-polar groups while infrared spectroscopy is most appropriated for asymmetric vibrations of polar groups [52].

Sir C. V. Raman and K. S. Krishnan were the first who succeeded in demonstrating the inelastic scattering of light by a fluid in 1928 [53]. They observed that the wavelength of the radiation scattered by certain molecules was differing from the wavelength of the incident beam of radiation and that the wavelength shift depended on the chemical structure of the molecules scattering the radiation [49]. Based on this discovery, Raman was awarded with a Nobel Prize in physics in 1930 [53]. Before the advent of laser sources, Raman spectroscopy was very limited and used only by specialized laboratories. With the development of laser sources in 1960s, Raman spectroscopy became more common and the capabilities and availability of instrumentation progressed rapidly in the 1980s and 1990s. Raman is now a common and popular technique due to the commercial availability of high performance instrumentation at moderate costs [49,53].

Raman and infrared spectroscopy are complementary vibrational techniques that differ fundamentally based on the interaction between radiation and molecules to probe vibrational states. A molecular vibration is infrared active if there is a change in the dipole moment of the molecule associated with the absorption of radiation [49]. Raman spectroscopy, on the other hand, is related to the polarizability of molecules and vibrational modes of molecules are Raman active only if they are associated with a change in polarizability of a molecule [49,52]. Polarizability consists of a momentary distortion of the electrons distributed around a bond in a molecule generated by an oscillating external electric field from the incident radiation, and its relaxation returning to the normal state by reemitting radiation [49,52], it's demonstrated in figure 3.

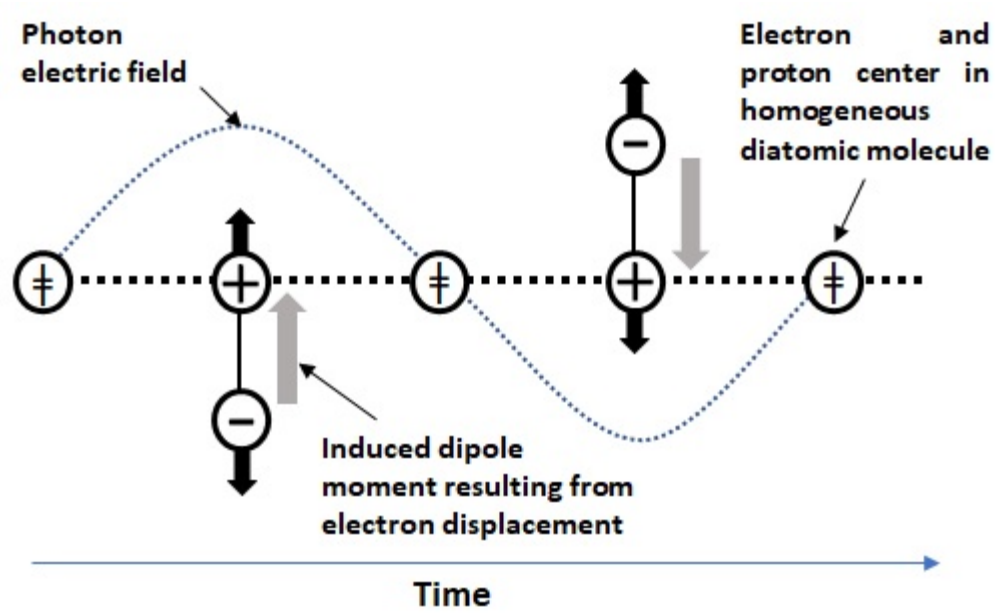

Figure 3. Demonstration of polarizability process in a homogeneous diatomic molecule. The oscillating electric field of the incident radiation induces a momentary induced dipole moment.

Raman spectroscopy is performed by irradiating the sample with a nearly monochromatic laser source generally in the UV to NIR spectral region with an excitation wavelength of higher energy than the absorptions bands of associated vibrational modes. Absorption of photons at this wavelength promotes the molecule to a higher virtual energy level *j* [49]. The radiation will induce a change in the polarizability of the molecule, and then, relaxation occurs generating a fraction of radiation beams scattered in all directions. The detection of this scattered radiation is carried out under some angle (generally, 180°) with a suitable spectrometer [49].

There are three different modes of scattering radiation, one is classified as Rayleigh process and the two others are Raman processes, they are described in figure 4. In the Rayleigh process no energy is lost, which is referred to as elastic scattering where a photon of the same energy as the excitation photon, $h\nu_{ex}$, is emitted. In Raman processes, inelastic scattering occurs where, the scattered photon results in a transition from the ground state to the first excited vibration state or vice-versa. Anti-Stokes Raman scatter results when the scattered photon is higher energy than the exciting photon, $h.(\nu_{ex}+\nu_v)$, and Stokes Raman scatter results when the scattered photon is lower energy than the exciting photon, $h(\nu-\nu_v)$.

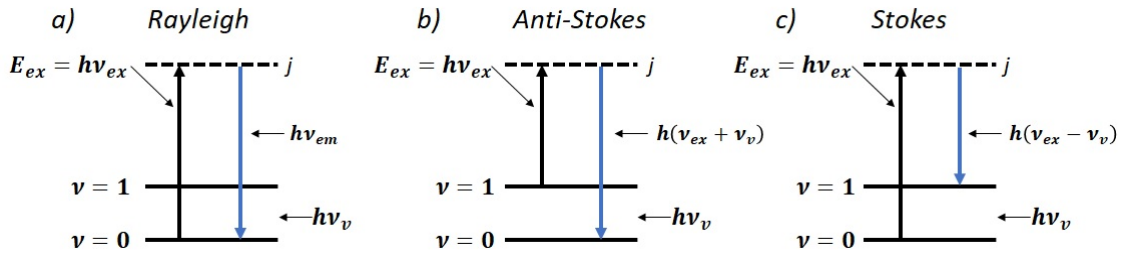

Figure 4. Schematic illustration of Rayleigh and Raman scattering modes occurring by excitation and relaxation of a molecule between the virtual energy level *j* and the vibrational energy levels ($\nu = 0$ and $\nu = 1$).

Figure 5 is representative of a simple Raman spectrum at a typical ambient temperature. The intensity of Stokes scatter is much higher than that of anti-Stokes scatter because the majority of molecules are in the ground state and, consequently, absorb energy corresponding to the induced vibrational mode in the molecule. As the temperature is increased, the ratio of Anti-Stokes to Stokes intensities increases since there will be more molecules present in the excited vibrational state [49,52]. Because the intensity of Stokes Raman scatter is generally much stronger than anti-Stokes, instruments are often designed to measure only the Stokes Raman spectrum [49]. Raman intensity is related to several parameters, the number of scattering molecules being among them. Consequently, if sample and measurement conditions are consistent, band intensity is related to analyte concentration and Raman spectroscopy can be used to perform quantitative analysis [52].

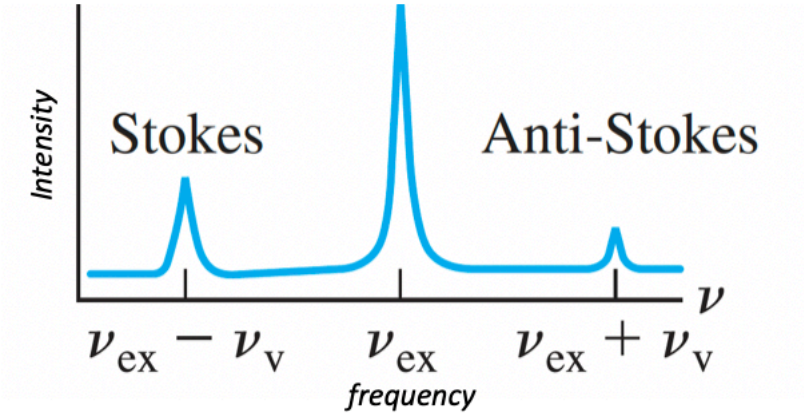

Figure 5. Typical Raman spectrum with Stokes presenting higher intensity than Anti-Stokes because the majority of molecules are on the ground state instead of the first vibrational level. (figure adapted from D.A. Skoog, Principles of Instrumental Analysis, 7th edition, 2016, p. 438).

3.2. Chemometrics

3.2.1. Data organization

NIR and Raman data are collected in such way that a sample *i* is represented by a row vector, where, each element $x_{i,j}$ of this vector is named variable, and those variables represent the absorbance (for NIR) or the intensity (for Raman) obtained for specific energy values, which, in this work were represented by wavelengths or wavenumbers. Once that the spectra of all samples are acquired, a matrix **X** of data is generated by concatenating all spectra in row direction as exemplified in figure 6.

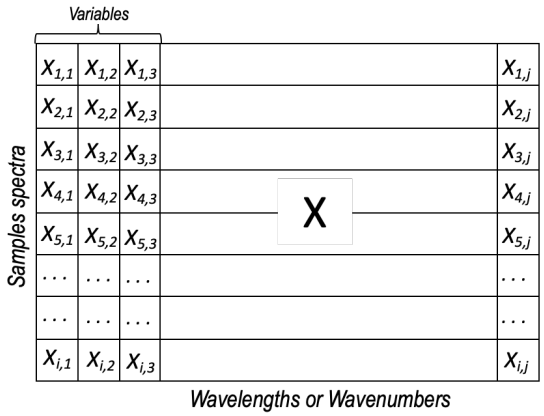

Figure 6. Mathematical representation of spectra. Each row corresponds to a sample *i* and each column *j* represents a different energy value related to wavelengths or wavenumbers, the element $x_{i,j}$ is a variable containing a value for the absorbance or intensity.

3.2.2. Preprocessing

NIR and Raman spectra usually present some systematic errors that can be generated by, for example, fluctuations of the system (source, detector), physical changes of positioning the sample, among other factors. Those systematic errors result in sample spectra which differ one to another, being those differences noted by the presence of background, noise, scattering.

In order to model properly the relationship between spectra and a property of interest, excluding or minimizing the influence of systematic errors is necessary. This can be reached by applying mathematical preprocessing techniques to the data prior to the modelling process. The preprocessing techniques which have been used in this work are described below.

a) Mean centering

Mean center (MC) is a very simple preprocessing technique which consists basically of subtracting the mean column $\bar{x_j}$ from each value of variable *j* generating the new value $x'_{i,j}$ as it is demonstrated in equation 3.3. The use of this technique makes easier to interpret the data, once that now the mean of the data is zero for all variables.

$$x'_{i,j} = x_{i,j} - \bar{x}_j \qquad \textbf{(3.3)}$$

b) Autoscaling

The procedure of autoscaling gives to all variables the same weight, its use is important in case that different variables are measured under very different ranges [54]. The calculation is showed in equation 3.4, it consists of mean centering the data, and then, divide each variable by its standard deviation s_j, and $x'_{i,j}$ is the new value.

$$x'_{i,j} = \frac{\left(x_{i,j} - \bar{x}_j\right)}{s_j} \qquad \textbf{(3.4)}$$

c) Normalization

Sometimes, spectroscopic observations are not comparable one to each other due to the presence of some undesired effects resulted from systematic biases that are usually generated by, for example, source or detector fluctuations, physical positioning of the sample. Normalize can be applied on the data in order to overcome those effects and provide equal importance to all samples by dividing the variables by a scaling factor. In this work the scaling factor w_i was calculated by area as described in equation 3.5, and then, the normalized variable $x'_{i,j}$ is obtained by dividing the original variable by the scaling factor, equation 3.6.

$$w_i = \sum_{j=1} |x_{i,j}| \qquad \textbf{(3.5)}$$

$$x'_{i,j} = x_i {w_i}^{-1} \qquad \textbf{(3.6)}$$

d) Multiplicative Scatter Correction

Multiplicative Scatter Correction (MSC) is performed in order to correct light scattering. MSC consists basically of two main characteristics, the first one assumes that a spectrum $\mathbf{x}_\mathrm{i}^\mathrm{T}$ is formed by light diffusion $\mathbf{d}_\mathrm{i}^\mathrm{T}$ and chemical absorbances $\mathbf{c}_\mathrm{i}^\mathrm{T}$ contributions as demonstrated in equation 3.7 where the superscript T indicates transposition; and the other characteristic assumes that the coefficients for the light diffusion contribution is the same for all samples, and then, it can be fitted by least squares by using a reference spectrum (usually, the average) [55].

The calculation is carried out by firstly averaging the spectra in order to obtain the reference spectrum, $\bar{x}$, and lately, the least squares are computed between the reference spectrum and the sample spectrum x_i, where the intercept *b* and the slope *a* are obtained, and finally, the MSC preprocessed spectra $x'_{i,j}$ is obtained by using the equation 3.8 [55,56].

$$\mathbf{x}_\mathrm{i}^\mathrm{T} = \mathbf{d}_\mathrm{i}^\mathrm{T} + \mathbf{c}_\mathrm{i}^\mathrm{T} \qquad \textbf{(3.7)}$$

$$x'_{i,j} = \left(x_{i,j} - b\right)/a \qquad \textbf{(3.8)}$$

e) Standard Normal Variate

Standard Normal Variate (SNV) is quite related to MSC, where, the main purpose of its use is related to correct light scattering effects. Its calculation is very simple, each spectrum is centered and then scaled by using the standard deviation s_i of the absorbance values for sample *i*. Equation 3.9 represents how the preprocessed matrix $x'_{i,j}$ is obtained after SNV is applied [57].

$$x'_{i,j} = \frac{\left(x_{i,j} - \bar{x}_i\right)}{s_i} \qquad \textbf{(3.9)}$$

f) Smoothing and Derivatives

Derivatives are preprocessing tools that are usually applied in spectroscopic data in order to remove the baseline effect among the samples, however, derivatives have a drawback of de-emphasize lower frequencies and emphasize high frequencies, it means that derivatives can

accentuate noises [58]. In order to overcome this effect of noise accentuation, Savitzky-Golay derivatives algorithm smooths the data prior to the derivatization. The smoothing procedure consists of fitting polynomials with a fixed odd windows width around of the spectrum resulting in noise reduction [57]. The windows width should be chosen according to the wanted noise reduction [57].

g) Baseline (Whittaker filter)

It consists of a preprocessing technique used for correcting the baseline effects on spectra using Whittaker smoother. This is calculated according to equation 3.10, where a generalized least squares function is performed to create a vector $\mathbf{z}_i^t$ that is smooth, and faithful to the spectrum $\mathbf{x}_i^t$ [59,60].

$$S = \sum_i \mathbf{w}_i^T \left(\mathbf{x}_i^T - \mathbf{z}_i^T\right)^2 + \lambda \sum_i \left(\Delta^2 \mathbf{z}_i^T\right)^2 \qquad \textbf{(3.10)}$$

The first term in S is related to the fitting to the data, while the second term refers to the non-smooth behavior of $\mathbf{z}_i^t$. The balance between the two terms is adjusted by the parameter λ. $\mathbf{w}_i^T$ is the vector of weights [60].

Usually, the sign of the residuals $\mathbf{x}_i^t - \mathbf{z}_i^T$ does not matter, with positive and negative residuals having the same weights. However, in this method it was observed that useful results are obtained when much more weight is given to the negative results. The parameter *p* to compute the weights is then introduced obeying to the following rules: $\mathbf{w}_i^T = p$ if $\mathbf{x}_i^t > \mathbf{z}_i^T$ and $\mathbf{w}_i^T = 1 - p$ in any other way [60].

To calculate this method of baseline correction the user needs to determine two parameters: *p* for asymmetry and λ for the smoothness. Both of them are adjusted according to the data at hand. Generally, values of $0.001 \leq p \leq 0.1$ and $10^2 \leq \lambda \leq 10^9$ appear to be good choices [60].

3.2.3. Principal Component Analysis

NIR and Raman spectroscopies consist of modern instrumentations which provide data composed by thousands of measured variables as result. Sometimes, these data can be overloaded, it means that some of the measured variables do not contain any useful information, thus, it is necessary to use some tool which can compress the data maintaining its essential information. Principal Component Analysis (PCA) is a largely used tool for that purpose.

PCA is an exploratory method of analysis which can be used for different purposes, such as, data reduction, outlier detection, or classification. It is a statistical method which tries to use linear combinations called principal components (PCs) to explain the variance of a data [55]. It consists of a decomposition of the original data matrix **X** of *m* rows (samples or objects) and *n* columns (variables), into a sum of *k* $\mathbf{t}_i$ and $\mathbf{p}_i^T$ [56,57,61]. The decomposition of matrix **X** is represented in equation 3.11 and figure 7.

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \cdots + \mathbf{t}_k\mathbf{p}_k^{\mathrm{T}} + \mathbf{E} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad \textbf{(3.11)}$$

Where, *k* is the number of PCs selected for the model, the superscript T denotes transposition, the product $\mathbf{t}_k\mathbf{p}_k^{\mathrm{T}}$ is the *k*-th principal component; $\mathbf{t}_k$ vectors are the *scores* which explains how the samples are related to each other; $\mathbf{p}_k^{\mathrm{T}}$ vectors are the *loadings* which can be related to the information on how the variables relate to each other; **E** is the residual matrix, it means, the unexplained part of **X** [56,57,61].

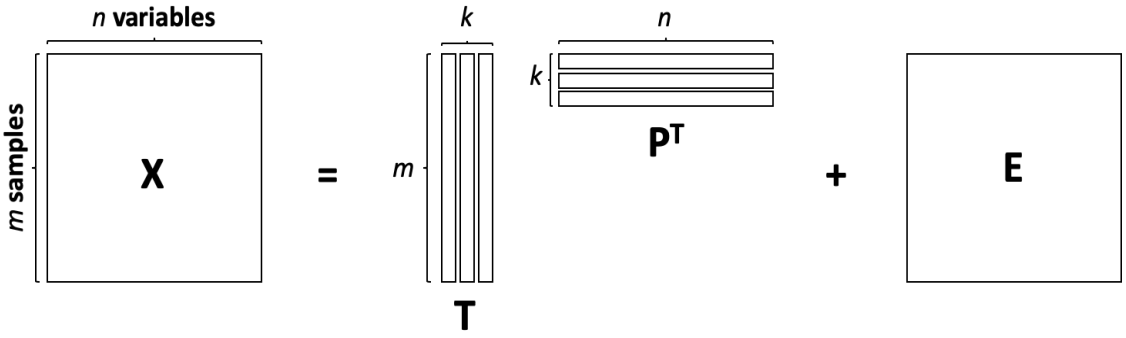

Figure 7. Representation of PCA decomposition of matrix X into *k* principal components which are composed by scores **T**, loadings $\mathbf{P}^{\mathrm{T}}$, and the residuals **E**.

Each PC is calculated in such a way to preserve the maximum variance among the objects; hence, the PCs are ranked in descending order according to the variance. It means that PC1 contains the maximum variance, PC2 is calculated using the residual matrix obtained after subtracting PC1 from **X** resulting in a PC orthogonal to PC1, it represents the second biggest variance; the remaining PCs are obtained in the same way as PC2, always using the residual obtained after subtracting the sum of the previous PCs from **X**, generating smaller variances than the previous PCs [54,62].

One very important step when performing PCA is to determine the adequate number of PCs to be used. If only a few PCs are selected to build the PCA model, part of the information described by the variables will not be taken into account by the model and information is lost resulting in a underfitting of the model. In other hand, when too many PCs are selected for the PCA model, residual (noise) are being included and consequently, the model will not be robust enough when applied to a new data, it denotes an overfitting of the model. Scree plot is a common method employed to determine the ideal number of PCs to be used in a PCA model. This consists of plotting the variance *vs.* principal component number, the biggest variance values are concentrated on the first principal components, so, scree plot shows a steep descent region followed by a flat region, this principal component which divides those two regions represents the ideal number of PCs to be used by the PCA model [57]. Another common method is the cross-validation leave one out, which a sample from the matrix **X** is left out and the PCA model is calculated using the remaining samples, then the left-out sample is used to estimate the performance of the model, the procedure is repeated until all samples have been left out once [63].

3.2.4. Soft Independent Modelling of Class Analogy

Defined classes cannot be differentiated by using only PCA by itself as the class information is not used during the model calculation and PCA just describes the overall variation in

the data [64]. In contrast of that, Soft Independent Modelling of Class Analogy better known by its acronym SIMCA is a supervised method of classification. Supervised method means that the class information is used to obtain the classification model.

SIMCA works basically like this: the data from each class of the calibration set is taken and a completely independent PCA model is calculated for these data. This means that different preprocessing techniques and number of PCs can be used for the PCA model of *z* different classes. The appropriate number of PCs for each PCA model is determined using cross-validation with scree plot as a guideline to do not include PCs containing small variance that can be modelling noise. This procedure is repeated until a PCA model is obtained for each class, and these PCA models define the SIMCA model.

The classification of a new unknown sample $\mathbf{x}_\mathrm{u}$ is determined based on how close this sample is from the classes. This is carried out by firstly projecting this sample into the PCA models for each class. In order to do that, the scores $\mathbf{t}_j$ of $\mathbf{x}_\mathrm{u}$ in the PCA space are calculated as shown in equation 3.12, where, $\bar{\mathbf{x}}_\mathrm{j}$ is the group center of class *j*. Then the estimation $\hat{\mathbf{x}}_\mathrm{u}$ of $\mathbf{x}_\mathrm{u}$ is obtained by performing a back-transformation of these scores to the original space as represented in equation 3.13 [55].

$$\mathbf{t}_j = (\mathbf{x}_\mathrm{u} - \bar{\mathbf{x}}_\mathrm{j}) \cdot \mathbf{p}_j \qquad \textbf{(3.12)}$$

$$\hat{\mathbf{x}}_\mathrm{u} = \mathbf{t}_j \cdot \mathbf{p}_j^\mathrm{T} + \bar{\mathbf{x}}_\mathrm{u} \qquad \textbf{(3.13)}$$

The next step consists of calculating the orthogonal distance between the new sample and the PCA space of class *j* using equation 3.14.

$$\mathrm{OD}_u^j = \|\mathbf{x}_\mathrm{u} - \bar{\mathbf{x}}_\mathrm{u}\| \quad \text{for } j = 1, \dots, z \qquad \textbf{(3.14)}$$

Finally, sample $\mathbf{x}_\mathrm{u}$ is classified based on an *F*-test $\left(s_u / s_j\right)^2$. The terms s_u^2 and s_j^2 are calculated according to equations 3.15 and 3.16, respectively. Where, OD_i^j is the orthogonal distance between the *i*-th sample of class *j* and the PCA model of class *j*, *p* is the number of variables, k_j is the number of PCs retained by PCA model of class *j*, and n_j is the number of samples in class *j* [55].

$$s_u^2 = \frac{\left(\mathrm{OD}_u^j\right)^2}{p - k_j} \qquad \textbf{(3.15)}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j}\left(\mathrm{OD}_i^j\right)^2}{(p - k_j)(n_j - k_j - 1)} \qquad \textbf{(3.16)}$$

If the obtained value for *F* is smaller than the critical value, sample $\mathbf{x}_\mathrm{u}$ belongs to class *j*. However, SIMCA has the property of soft modelling. This property establishes that a sample $\mathbf{x}_\mathrm{u}$ can be classified as belonging to more than one class or as not belonging to any class [55,57].

3.2.5. Partial Least Squares

Partial Least Squares (PLS) regression is one of the most commonly used method for multivariate calibration applied for quantitative analysis [55,57]. It consists of finding a relationship between independent variables in a matrix **X** (NIR or Raman spectra) and dependent variables of a vector **y** (Gasoline properties) by using linear combinations in order to obtain latent variables (LV) which are calculated in such a way to obtain the maximum covariance between **X** and **y**.

PLS calculations occurs by decomposing matrix **X** and vector **y** into a product of scores (**T** and **U**) and loadings ($\mathbf{P}^T$ and $\mathbf{Q}^T$) as respectively represented by equations 3.17 and 3.18, and then, an inner relation is established by connecting **X**-scores with **y**-scores giving them information about each other; lately, weights are introduced in order to obtain orthogonal **X** scores. $\mathbf{E}_x$ and $\mathbf{E}_y$ are the residual for matrix **X** and vector **y**, respectively [57,62].

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{\mathrm{x}} \qquad \textbf{(3.17)}$$

$$\mathbf{y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{E}_{\mathrm{y}} \qquad \textbf{(3.18)}$$

One important step when using PLS regression is related to the ideal number of latent variables to be included in the model. The same problems related to underfitting and overfitting which were previously explained for PCA can also occurs for PLS, it can be minimized by using a method called cross-validation (CV). Cross-validation is made by splitting the calibration (CAL) samples into CV groups (in such a way that all CAL samples have composed at least one CV group), the calibration of the model is carried out with the remaining calibration samples, then the CV groups are predicted by this model one at a time and the model performance can be estimated, It allows the user to select the ideal number of LV's to be used [65,66]. There are many different CV methods such as: leave-one-out, contiguous block, venetian blinds, etc. Venetian Blinds [65] was the chosen method for this work. It divides the CAL samples in *n* CV groups containing *m/n* samples each, where *m* is the total number of samples. Figure 8 shows how the CAL samples are split in CV groups by venetian blinds approach, where each color represents samples that compose a CV group.

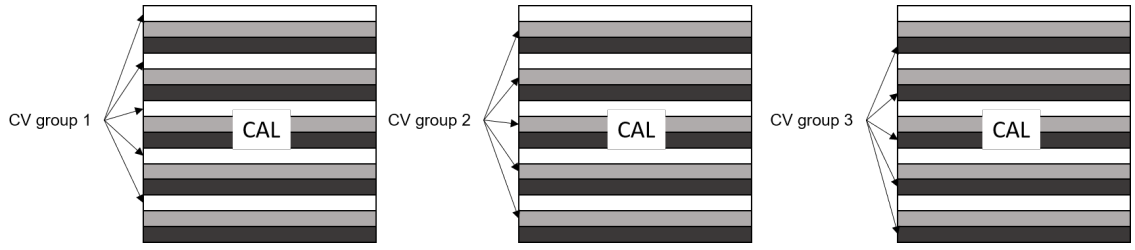

Figure 8. Illustration of cross-validation venetian blinds. Each row represents a sample and each color represents a CV group.

The PLS models were also evaluated using the response residual plots (predicted values vs. studentized residuals). These plots are an indication of the lack of fit of the predicted value of a sample. The studentized residuals have mean 0 and variance 1, and good PLS models are expected to have studentized residuals randomly distributed around the mean [67,68]. Samples with high studentized residual values consist of samples that could not be fitted by the PLS model, thus, being considered outliers. These outliers can be related to the presence of systematic errors or samples that were not well represented in the calibration model.

3.2.6. Figures of merit

Figures of merit consist of quality parameters which are used to compare the performance of different analytical methods; and it occurs by means of numerical indicators which are of easy and simple interpretation. The analyst can take into account the figures of merit and other factors related to the methodology (such as cost, time of analysis, possibility of automation, etc.) in order to choose the most appropriate analytical method before a specific application [58].

a) Qualitative analysis

In this work, the performance of classification was carried out using the following figures of merit: sensitivity (SEN), specificity (SPE), precision (PREC), and misclassification error (ME).

Misclassification error for each class is calculated as showed in equation 3.19, where, C_{cp} is the number of samples correctly predicted for a determined class, and C_{ref} is the real number of samples for this specific class.

$$\mathrm{ME} = \frac{\left(C_{ref} - C_{cp}\right)}{C_{ref}} \times 100\% \qquad \textbf{(3.19)}$$

Sensitivity represents the ability of the classification model in to correctly identify samples of determined class, it is calculated according to equation 3.20 [69].

$$\mathrm{SEN} = \frac{C_{cp}}{C_{ref}} \qquad \textbf{(3.20)}$$

Precision is related to the purity of a class, the ability of the classification model in avoiding wrong predictions for determined class, it is computed as represented in equation 3.21, where, C_p is the total number of samples predicted for a specific class [69].

$$\mathrm{PREC} = \frac{C_{cp}}{C_p} \qquad \textbf{(3.21)}$$

The specificity of a class is determined as the ability of the classification model in reject samples which do not belong to the class being analyzed, its calculation is carried out according to equation 3.22, C_{nb} is the number of samples not belonging to the class being analyzed which were not classified in this class, and N is the total number of samples [69].

$$\mathrm{SPEC} = \frac{C_{nb}}{\left(N - C_{ref}\right)} \qquad \textbf{(3.22)}$$

The results for both sensitivity, precision, and specificity will consist of values ranging from 0, where there is no class discrimination; and 1, when the classification model presents a perfect class discrimination [69].

b) Quantitative analysis

In order to assess the performance of PLS models obtained for the quantitative analysis, the following figures of merit were considered: the root-mean-square error (RMSE); the correlation coefficients (R^2); the average relative errors (ARE); and the elliptical joint confidence region (EJCR).

RMSE is obtained according to equation 3.23 and it can be used for both calibration, cross validation, and prediction sample sets, being RMSEC, RMSECV, and RMSEP their acronyms, respectively; $\hat{y}_i$ is the value predicted by the PLS model, y_i is the reference value and *N* is the total number of samples.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}(\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2}{\mathrm{N}}} \qquad \textbf{(3.23)}$$

The correlation coefficient R^2 is determined as showed in equation 3.24. It is also computed for calibration, cross validation and prediction sample sets which are represented by R_C^2, R_{CV}^2, and R_P^2, respectively. In equation 3.24, $\chi_{x,y}$ is the covariance of reference and predicted values; σ_x and σ_y are the standard deviations of the reference and the predicted values, respectively.

$$R^2 = \left(\frac{\chi_{x,y}}{\sigma_x \cdot \sigma_y}\right)^2 \qquad \textbf{(3.24)}$$

The calculation for estimating the average relative error is explained by equation 3.25.

$$ARE = \frac{\sum(|\hat{y}_i - y_i|/y_i)}{N} \times 100\% \qquad \textbf{(3.25)}$$

EJCR consists of a statistical tool used to assess the accuracy of the model. A linear regression between the reference values and predicted results is fitted, and then, the obtained values for intercept (b) and slope (a) are estimated by an ellipse with a confidence level of 95%, so, these estimative values are compared with their ideal theoretical values, 0 and 1, for intercept and slope, respectively. If the point (0,1) lies inside the confidence ellipse, it indicates that BIAS is absent and the model is accurate according to the reference method [70,71].

4. EXPERIMENTAL

4.1. Samples

All samples used in this study were transferred to amber glass bottles, sealed with Teflon tape and stored in the fridge at the temperature of -4 ºC until use. The samples were allowed to come up to ambient laboratory temperature (22-25 ºC) before spectroscopic analyses.

4.1.1. Qualitative analysis

Four gasoline sample sets were used in this study resulting in a total of 124 samples. The collection of samples is illustrated in the diagram presented in figure 9a.

The largest sample set consisted of 99 automotive commercial gasoline samples of 18 different brands which were purchased in 27 gas stations in 14 cities of 5 states in the eastern United States of America. The samples were acquired over a period of 5 months from July to November of 2018. These samples were collected with the objective of maximizing variability with respect to antiknock index (AKI), ethanol content (ranging from ethanol-free up to 15%) and brands.

Six samples composed the second sample set. These samples consisted of synthetically blended gasoline samples provided by the Environmental Protection Agency (Washington, D.C., United States).

The third set was composed of standard reference materials (SRM's) available from the National Institute of Standards and Technology (Gaithersburg, MD, United States). The following SRM's were obtained: ethanol in reference gasoline (SRM 2287), t-amyl-methyl ether in gasoline (SRM 2289), ethyl-t-butyl ether in gasoline (SRM 2291), reformulated gasoline 11% MTBE (SRM 2294), reformulated gasoline 15% MTBE, reformulated gasoline 13% ETBE (SRM 2296), reformulated gasoline 10% ethanol (SRM 2297), sulfur in gasoline-high octane (SRM 2298) and a gasoline blank provided in the SRM 2287 kit.

A total of 8 gasoline blends designed by the Fuels for Advanced Combustion Engines (FACE) [72] group were purchased from the Coordinating Research Council (Alpharetta, GA, United States).

4.1.2. Quantitative analysis

Some of the samples described above were selected for use in the quantitative analysis as well, among them: 6 EPA samples, 4 FACE samples, and 52 of the automotive commercial gasoline samples that were chosen based on their variability of AKI and ethanol content. The variability of those samples used for the quantitative analyses is showed in figure 9b.

From the 4 FACE samples, 4 new samples have been prepared by adding anhydrous ethyl alcohol (U.S.P., 200 Proof, Warner Graham Co., Cockeysville, MD) at 10% v/v. These were prepared because the properties of these 4 FACE fuels were also characterized as blends with different amounts of ethanol.

A total of 100 samples were acquired from the Summer 2018 North American Fuel Survey (NAFS) conducted by the Alliance of Automobile Manufacturers (Washington, D.C., United States). This is a subset of 500 commercial gasoline samples collected for this biannual survey. The NAFS summer 2018 survey involved the sampling of commercial automotive gasoline samples labeled as regular and premium grades (mid-grade samples are not collected) from 44 cities across the US (29), Canada (8) and Mexico (7). Gasoline with indicated ethanol content above 10% are not sampled in these surveys. The 100 samples utilized in the present work included fuels from 18 states of the United States plus a few samples from Mexico and Canada.

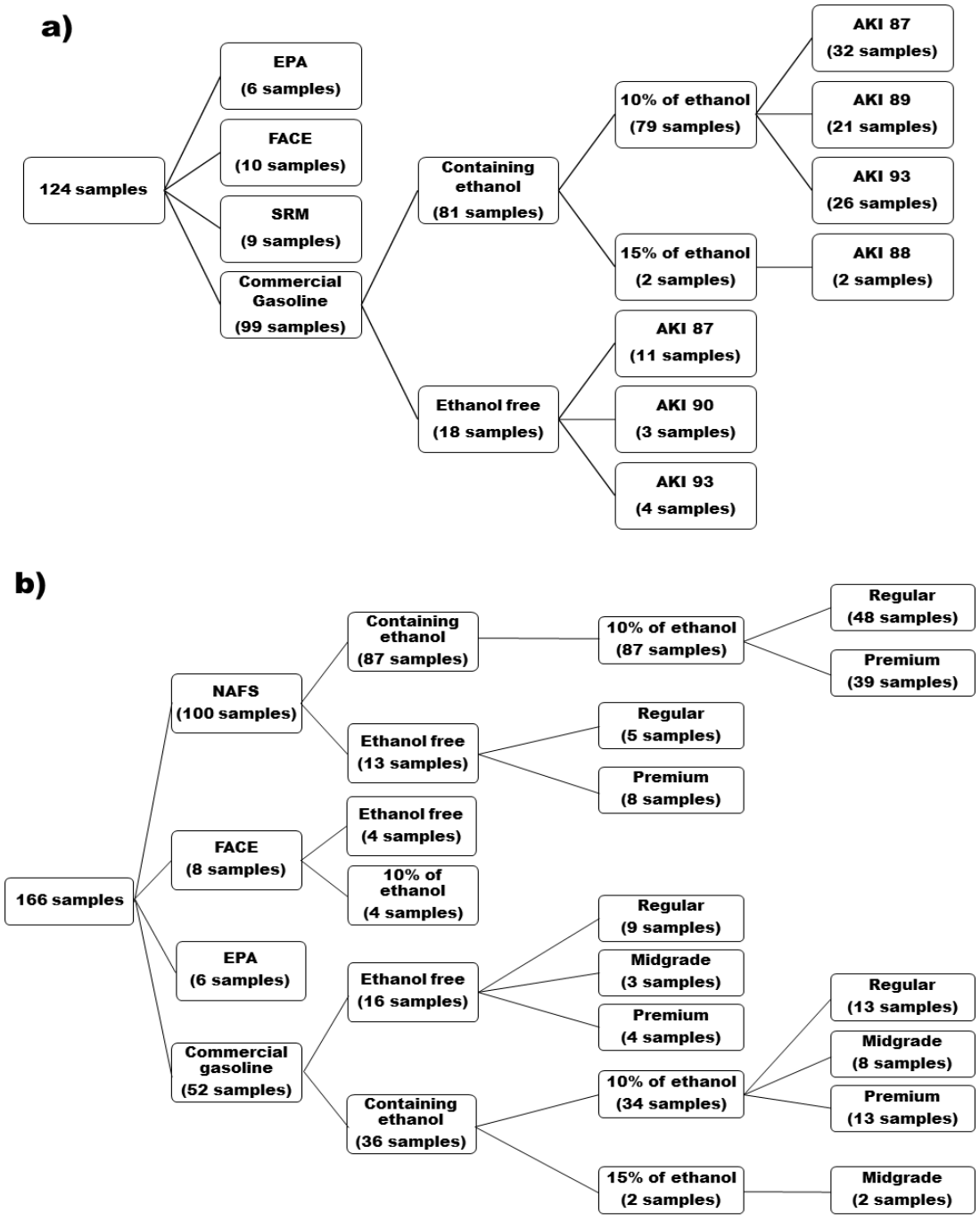

Figure 9. Diagram representing the distribution of gasoline samples used for a) qualitative analysis b) quantitative analysis.

4.2. Reference values

4.2.1. Qualitative analysis

For the commercial samples, the posted octane rating (AKI) and maximum ethanol content available on the gasoline pumps at the selected gas stations where samples were purchased were used as reference values for qualitative evaluation of the data. For the remaining samples (FACE, EPA, and SRMs), reference values available on reports or certificates were used.

4.2.2. Quantitative analysis

a) Benzene, Aromatics, Olefins and Paraffins:

A proton nuclear magnetic resonance method (^{1}H NMR) based on the integration of selected regions of the NMR spectrum and the average group molecular weights was used as a reference method for providing quantitative results for aromatics, olefins, paraffins and benzene expressed in percent by volume [73,74].

Deuterated chloroform (chloroform-d "100%", with 0.05% v/v tetramethylsilane) acquired from Cambridge Isotope Laboratories (Tewksbury, MA, United States) was used for the NMR sample preparation. Samples were prepared by diluting 50 µL of gasoline in 500 µL of $CDCl_3$ in a 2.0 mL HPLC vial glass. Gasoline samples were dispensed below the surface of the chloroform using a positive displacement glass bore micro-pipette to minimize losses from evaporation. The samples were then transferred to NMR tubes and sealed with Teflon caps to minimize loss of volatile compounds.

High-field NMR spectra were acquired on a Bruker Avance II 600 MHz spectrometer (Fällanden, Switzerland) equipped with a room temperature broadband inverse (BBI) probe. The spectra were acquired with the following parameters: sample temperature of 24.85 °C (298 K), 12000 Hz spectral width, 131072 complex data points, 90° pulse angle, 5s acquisition time, 25 s recycle delay, 32 scans, and a total acquisition time of approximately 17 minutes. The spectra were Fourier transformed followed by phase correction and baseline correction. The chemical shift axis was referenced to the tetramethylsilane (TMS) peak set to 0 ppm. The quantitative results for hydrocarbon class, oxygenate and benzene were calculated based on integration of the regions of interest.

The first step performed in the quantitative NMR analysis is the integration of the ^{1}H NMR spectral regions presented in Table 1. The integral values were then input into the sets of equations below. In these equations, V_x is the relative partial volume of compound class X; N_x is proportional to the number of molecules of X; MW_x is related to the mean molecular weight of compound class X; and ρ_x which represents the mean density of compound class X. After evaluating the results for each compound class, the concentrations in percent by volume (vol%) were later determined by using equation 4.1.

$$\mathrm{vol_x\%} = 100 * \mathrm{v_x}/(\mathrm{v_B} + \mathrm{v_{Ar}} + \mathrm{v_O} + \mathrm{v_P} + \mathrm{v_{Et}} + \mathrm{v_M}), \qquad \text{x = B, Ar, O, P, Et, M} \qquad \textbf{(4.1)}$$

In order to optimize the procedures of chemical shift correction related to the TMS peak, and obtaining the gasoline composition results, MATLAB functions were developed. It represented a good time saving for computing the results of all samples. The acquisition and adjust of phase for the High-field NMR spectra were performed using TOPSPIN 4.0.6 software developed by Brucker Biospin Corporation (Billerica, MA, United States).

Table 1. Assignment of regions to be integrated

Spectral Region	Chemical shift intervals (ppm)	Substance class
Ar	8.00 – 6.70	Aromatics
B	7.40 – 7.30	Benzene
C	7.27 – 7.26	Chloroform
O_1	6.00 – 5.75	
O_2	5.75 – 5.25	
O_3	5.25 – 5.05	Olefins
O_4	5.05 – 4.80	
O_5	4.80 – 4.60	
M	3.30 – 3.10	MTBE
A_1	3.00 – 2.80	
A_2	2.75 – 2.50	Aromatics
A_3	2.50 – 2.15	
O_e	2.10 – 1.85	Olefins
P	1.85 – 0.50	Paraffins
Et	4.30 – 3.30	Ethanol

i) Benzene

$N_B = B/6, MW_B = 78.1\ g.mol^{-1}, \rho_B = 879.0\ Kg.m^{-3}$

$V_B = (N_B \times MW_B)/\rho_B$

B is the integral value for the region assigned for benzene in table 1.

ii) Aromatics

$N_{Ar} = \left(A_r - C + A_1 + \frac{A_2}{2} + \frac{A_3}{3}\right)/6, \qquad \rho_{Ar} = 868.0\ Kg.m^{-3}$

$p_h = (A_r - C)/N_{Ar}, \qquad p_e = A_2/2N_{Ar}, \qquad p_m = A_3/3N_{Ar}, \quad p_p = A_1/N_{Ar}$

$MW_{Ar} = 72 + p_H + 15 * p_M + 29 * p_E + 43 * p_p$

$V_{Ar} = (N_{Ar} \times MW_{Ar})/\rho_{Ar}$

For aromatics composition, the average number of aromatic H atoms, methyl, ethyl and isopropyl substituents is determined by the parameters p_h, p_e, and p_m in order to calculate the mean molecular weight.

iii) Olefins

$$N_O = \frac{1}{2}O_1 + \frac{1}{4}O_4 + \frac{1}{2}O_2 + O_3 + \frac{1}{2}O_5 \qquad q = \frac{1}{N_O}\left(\frac{1}{2}O_1 + \frac{1}{4}O_4 + O_2 + 3O_3 + O_5\right)$$

$$q_e = \frac{1}{2} * \frac{O_E}{N_O}, \qquad q_m = q - q_e, \qquad q_h = 4 - q_m - q_e$$

$$MW_O = 24 + q_h + 15q_m + 29q_e, \qquad \rho_O = (1.775MW_O + 537)\ \mathrm{Kg.m^{-3}}$$

$$V_O = (N_O \times MW_O)/\rho_O$$

The parameters q_h, q_e, and q_m, are related to the mean number of olefinic hydrogens and ethyl and methyl substituents, respectively, and q is the average number of substituents per molecule.

iv) Paraffins

$$I_P = P - \left(6A_1 + \frac{3}{2}A_2\right) - \frac{3}{2}O_e\left(1 + \frac{q_m}{q_e}\right) - \frac{3}{2}Et - 3M$$

$$H_P = 15, \qquad MW_P = 93.0\ \mathrm{g.mol^{-1}}, \qquad \rho_P = 668.0\ \mathrm{kg.m^{-3}}$$

$$V_P = (I_P * MW_P)/(H_P * \rho_P)$$

From the paraffin integration, some regions were subtracted. Those regions are related to the contributions of the ethyl and isopropyl substituents of aromatics, the ethyl and methyl substituents of olefins, the tert-butyl group of MTBEs, and the methyl group of ethanol. H_p is the average number of hydrogen atoms per molecule.

v) Ethanol and MTBE

$$N_M = M/3, \qquad MW_M = 88.2\ \mathrm{g.mol^{-1}} \qquad , \rho_M = 741.0\ \mathrm{Kg.m^{-3}}$$

$$V_M = (N_M \times MW_M)/\rho_M$$

$$N_{Et} = Et/2, \qquad MW_{Et} = 46.0\ \mathrm{g.mol^{-1}}, \qquad \rho_{Et} = 791.5\ \mathrm{Kg.m^{-3}}$$

$$V_{Et} = (N_{Et} \times MW_{Et})/\rho_{Et}$$

While specific models were not developed for predicting ethanol and MTBE content, these values are needed to determine the concentrations for the other compound classes as well as benzene.

b) Relative Density, Research Octane Number, Motor Octane Number, Antiknock index:

The values for the gasoline parameters described in this section were obtained from the final report of the Summer 2018 North American Fuel Survey conducted by the Alliance of Automobile Manufacturers. The methods of analyses employed were the standard methods established by the American Society for Tests and Materials (ASTM).

i) Relative Density

The relative density is the ratio between the density of a substance at a fixed temperature and the density of a reference material at a stated temperature, in this case, water was used as reference material. The standard method for determining the relative density is described by the ASTM D-4052 [10]. It consists basically of adding a small volume (~ 0.7 mL) of gasoline into the previously cleaned and dried tube of the instrument; and then, waiting until the instrument presents a steady reading of four significant figures for density and five for temperature, it indicates that temperature equilibrium was reached and the value can be recorded as the final result.

ii) Research Octane Number

Research Octane Number (RON) is a test carried out in order to determine the fuel resistance to autoignition under mild conditions. The test was performed according to the standard method ASTM D-2699 [7] in which the result is obtained by comparison of the knock intensities for the test gasoline with the one obtained for a primary reference fuel blend at the same conditions using a standardized Cooperative Fuel Research (CFR) engine.

iii) Motor Octane Number

Motor octane number (MON) test is performed in a very similar way to the RON test. However, in this case, the test is performed under more severe conditions. The method used to determine MON for these samples was ASTM D-2700 [8].

iv) Antiknock Index

Antiknock index (AKI) represents the average value of RON and MON. This is the fuel octane rating that is generally posted on pumps at gas stations in the United States. This is also what fuel quality designations, such as regular, midgrade or premium is based on.

$$AKI = \frac{RON+MON}{2} \qquad \textbf{(4.2)}$$

4.3. Near infrared spectroscopy

4.3.1. Qualitative analysis

Near infrared transflectance spectra were recorded using a Bruker Vertex 70 FTIR spectrometer (Billerica, MA, United States) equipped with a Quartz-Tungsten-Halogen (QTH) radiation source combined with CaF_2 beam splitter, a thermo-electrically cooled InGaAs detector and a fiber optic probe. The samples were pipetted into 2 mL HPLC glass vials, and then, the vials were properly closed with their caps in order to avoid any losses of vapors. The spectra were collected at room temperature (22±2 °C) in the range between 4000-12000 cm^{-1} (800-2500 nm) in 64 scans with a resolution of 8 cm^{-1} taking 60 s per measurement, three measurements were

carried out for each sample. The spectra of the empty HPLC glass vial was used as background spectra.

4.3.2. Quantitative analysis

For the quantitative analysis the samples were analyzed using a 5 mm pathlength screw top quartz cuvettes. The spectra were obtained using the same spectrometer as in the qualitative analysis, however, the instrument was configured for the transmission mode instead of transflection. A Quantum Northwest temperature-controlled cuvette holder model t2 Sport (Liberty Lake, WA, United States) was coupled inside the instrument compartment, and the temperature controller was set to 22.5 °C. The other parameters (spectral range, number of scans, resolution) were maintained as the same ones used for the qualitative analysis.

4.4. Raman spectroscopy

4.4.1. Qualitative analysis

All measurements were carried out using a Bruker Vertex 70 FTIR with RAM II FT-Raman Module (Billerica, MA, United States) equipped with a 1.5 W Nd:Yag excitation laser, CaF_2 beam splitter, liquid-nitrogen cooled Ge-diode detector, and F/1.2 IR lens for 180° backscatter Raman measurements. The Raman laser power was set in the software at 700 mW with a fixed excitation wavelength radiation at 1064 nm, the spectra were recorded in the range between -50 cm^{-1} and 3500 cm^{-1} with a resolution of 4 cm^{-1} in 64 scans. The samples were analyzed in 2 mL HPLC glass vials, the vials were positioned with the radiation beam passing through its center. As demonstrated in figure 10, a piece of aluminum foil was placed and fixed behind of the sample in the sample holder in order to avoid spreading of radiation.

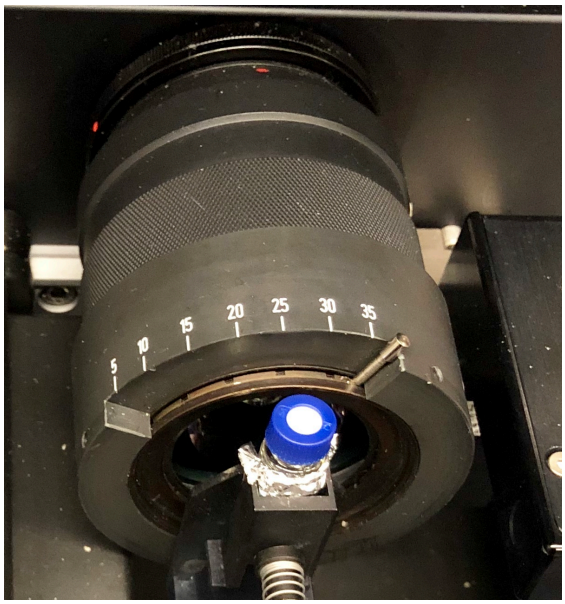

Figure 10. Illustration of how the samples were positioned for obtaining the Raman spectra. A piece of aluminum foil was placed and fixed behind of sample in the sample holder to avoid spreading of radiation.

4.4.2. Quantitative analysis

The 830 nm Raman measurements were conducted on a home-built Raman system based on illumination and collection via a low magnification 4X microscope objective (Olympus PLAN S-APO 4X, “Super Apochromat”) in a 180° backscatter geometry. Laser line blocking was achieved using two 25 mm 532 nm dichroic long-pass edge filters. The first filter (Semrock RazorEdge LP03-532RU-25, “U” grade, 186 cm-1 transition) was used as a beam steering mirror to introduce the laser onto the primary optical axis while the second (Semrock RazorEdge LP03-532RE-25, “E” grade, 90 cm-1 transition) was used to enhance laser line rejection. The remaining optical components consisted of several 1” protected silver beam steering mirrors (Thorlabs, PF10-03-P01) and an off-axis parabolic mirror (Thorlabs, MPD149-P01) to focus the collected Raman scatter onto the entrance slit of the spectrometer. The spectrometer was a 320 mm focal length, f/4.6, aberration corrected IsoPlane SCT-320 imaging spectrograph (Princeton Instruments, Acton, MA) equipped with a PIXIS 400BR eXcelon (Princeton Instruments) back-illuminated, deep-depletion, TE cooled (-70° C), 1340 × 400 pixel (20 μm pixels) CCD detector. The spectrometer was controlled by the LightField software (Princeton Instruments). Integration with LabView and synchronized acquisition with an XY stage was used to automate data collection from multiple samples.

The excitation laser used for this application was a frequency stabilized 830 nm diode laser (Innovative Photonic Solutions, Monmouth, NJ model I0830SR0100B) with nominal output power of 110 mW. A 600 grv/mm grating, blazed at 750 nm, was used with the center wavelength set to 905 nm for all measurements. This provided spectral coverage from approximately 843 nm to 965 nm (186 cm^{-1} to 1686 cm^{-1} Raman shift).

Samples were placed in 2 mL clear glass autosampler vials (ThermoFisher, SureStop C5000-592). Laser illumination and collection were through the bottom of the vials. A fixed integration time of 3s per spectrum was used.

4.5. Software and Computing

MATLAB R2018a software (MathWorks, Natick, MA, United States) was used for the mathematical analysis.

Spectral preprocessing, PCA, PLS and SIMCA were carried out using the PLS_toolbox 8.6.2 (Eigenvector Research, Inc., Manson, WA, United States).

5. RESULTS AND DISCUSSION

5.1. Qualitative analysis

5.1.1. NIR bands assignment

Figure 11a shows the NIR spectra obtained for the 124 gasoline samples used for the qualitative analysis. It is observed that the region located between 1670-1790 nm related to the first overtones of -CH, -CH_2 and -CH_3 bonds presents an absorbance saturation; also, the region related to the combination of C=C, -CH_3 and -CH_2 bonds placed between 2130-2500 nm is very noisy, so, it was decided to cut off those regions from the data set as showed in figure 11b. From the remaining spectra, two bands are more pronounced. The first band between 1100-1250nm corresponds to the second overtone of -CH functional group. The other band actually consists of two bands overlapping, one placed in the region between 1350-1500 nm which is related to the combination of -CH_3 and -CH_2 bonds and the other present about 1400-1670 nm corresponding to the first overtone of -OH bonds [75].

5.1.2. Raman bands assignment

The Raman spectra of 124 samples used for the qualitative analysis of gasoline is plotted in figure 11d. As the Raman detector has its sensitivity diminished for Raman shifts beyond 3200 cm^{-1}, the region 3150-3500 cm^{-1} does not show any relevant signal; another region situated in the range between 1700-2670 cm^{-1} also does not present any peak, so, those regions have been discarded in order to avoid including noise or other useless information in the Raman data set. The strong peak presented between 0-250 cm^{-1} consists of the Rayleigh scatter and low frequency Raman modes, thus being removed from the Raman data set as well. Figure 11e shows the remaining regions of the Raman spectra, it consists basically of two main regions, the region between 300-1800 cm^{-1} which it is mainly consisted of C-C skeletal vibrations and C-H deformations being known as fingerprint region [76], and the region placed between 2800-3100 cm^{-1} that is related to C-H stretching modes of ethanol and hydrocarbons [77].

5.1.3. Principal Component Analysis

Exploratory analysis by PCA was carried out on both the NIR and Raman data in order to observe the sample distribution and whether outliers were present or not. The NIR data was autoscaled and Raman data was mean centered prior to the PCA analysis. These preprocessing techniques were chosen in order to obtain the best separation among the clusters of gasoline based on ethanol content and antiknock index. The preprocessed spectra are shown in figures 11c and 11f for the NIR and Raman data, respectively. For the NIR data a total of 4 PCs resulting in 97.39% of cumulative variance were selected for the PCA model, and 3 PCs were chosen for the Raman data cumulating a total of 93.69% of variance.

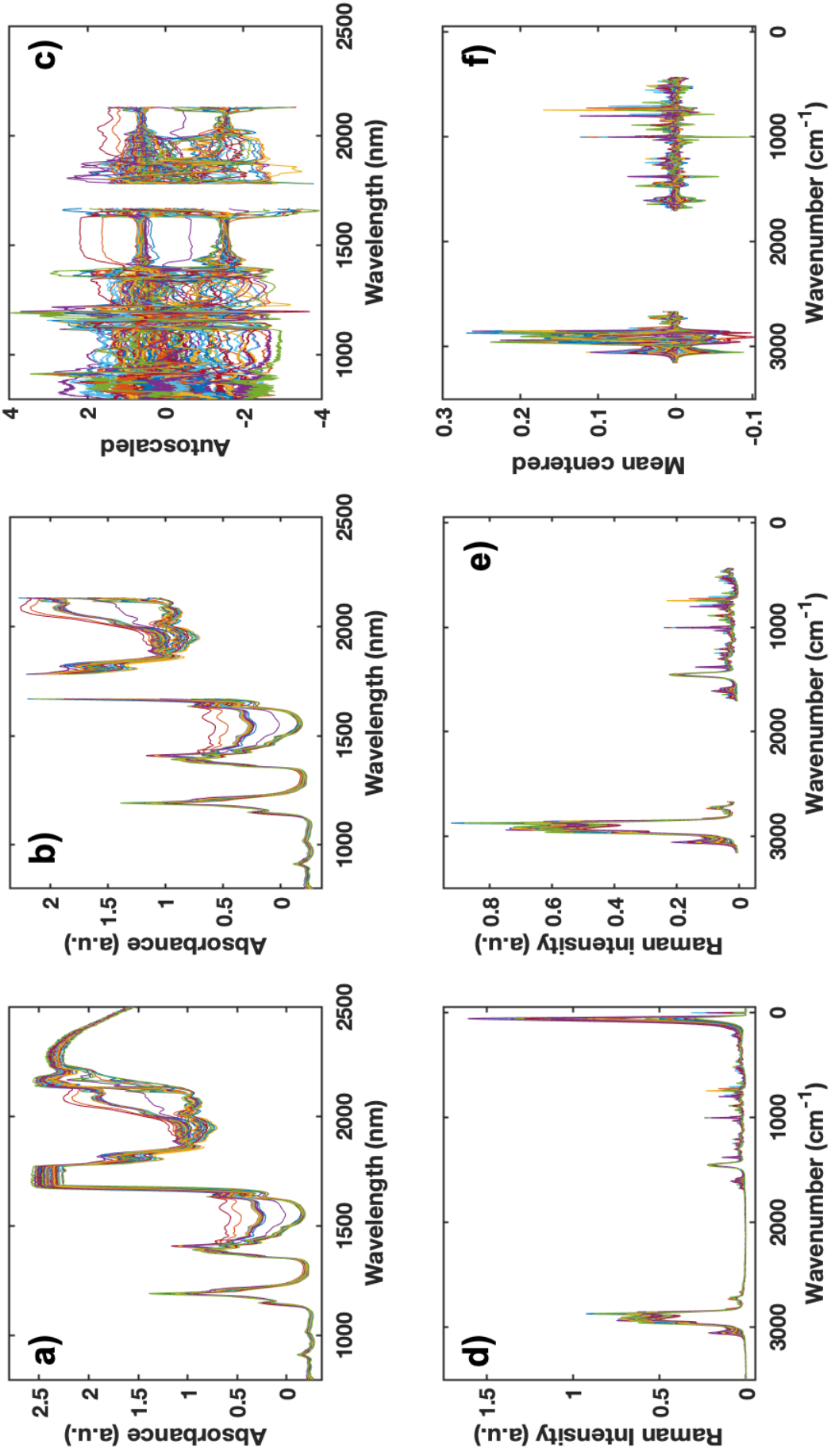

Figure 11. Original spectra obtained by a) NIR and d) Raman. Spectra obtained after excluding noisy and no-signal regions for b) NIR and e) Raman data. Preprocessed spectra for c) NIR and f) Raman

The outlier diagnostics were performed based on the plot of *Q* residuals against Hotelling's T^2, which is also known as an influence plot. These residuals are used as a tool to observe samples that could not be well-described by the PCA model [63]. The influence plot for the NIR and Raman data sets are shown in figures 12a and 12b, respectively. The FACE and SRM samples have higher residuals. This was not unexpected since these samples do not necessarily represent conventional gasoline samples like the commercial gasoline samples, once that some of them were synthetically prepared. Then, the only way to keep these samples for the analysis would be by including more samples analogously prepared, as it was not possible, it was preferred to do not include them in the data set. Only two SRMs (indicated in the figures by arrows) presented lower residuals and were kept in the data set.

After discarding the samples based on the outlier diagnostics, the PCA models for both NIR and Raman data were recalculated resulting in 4 PCs with 98.68% of cumulative variance for the NIR data, and 3 PCs with 98.63% of cumulative variance for the Raman data. Figures 12c-f displays combinations of PCA scores that provided the best class separation according to the ethanol content and AKI values. Figures 12c and 12d plots PC1 vs. PC2 scores for NIR and PC1 vs. PC3 scores for Raman, respectively. PC1 is highly correlated with ethanol content for both the NIR and Raman data sets. For the Raman data this PC alone can distinguish the ethanol content classes by itself while for the NIR data a contribution from PC2 is necessary. Good separations were obtained by both techniques. Figures 12e and 12f show the combination of PC scores for differentiate AKI classes by NIR and Raman, respectively.

The use of PCA, an unsupervised exploratory analysis technique, demonstrated that it was possible to distinguish gasoline sample classes based on ethanol content and AKI values using NIR and Raman data. However, as seen in these plots, class boundaries did exhibit some overlap. In addition to exploratory analysis by PCA, targeted classification of the gasoline samples was of interest. For that purpose, SIMCA, a supervised method of classification was applied to both NIR and Raman data.

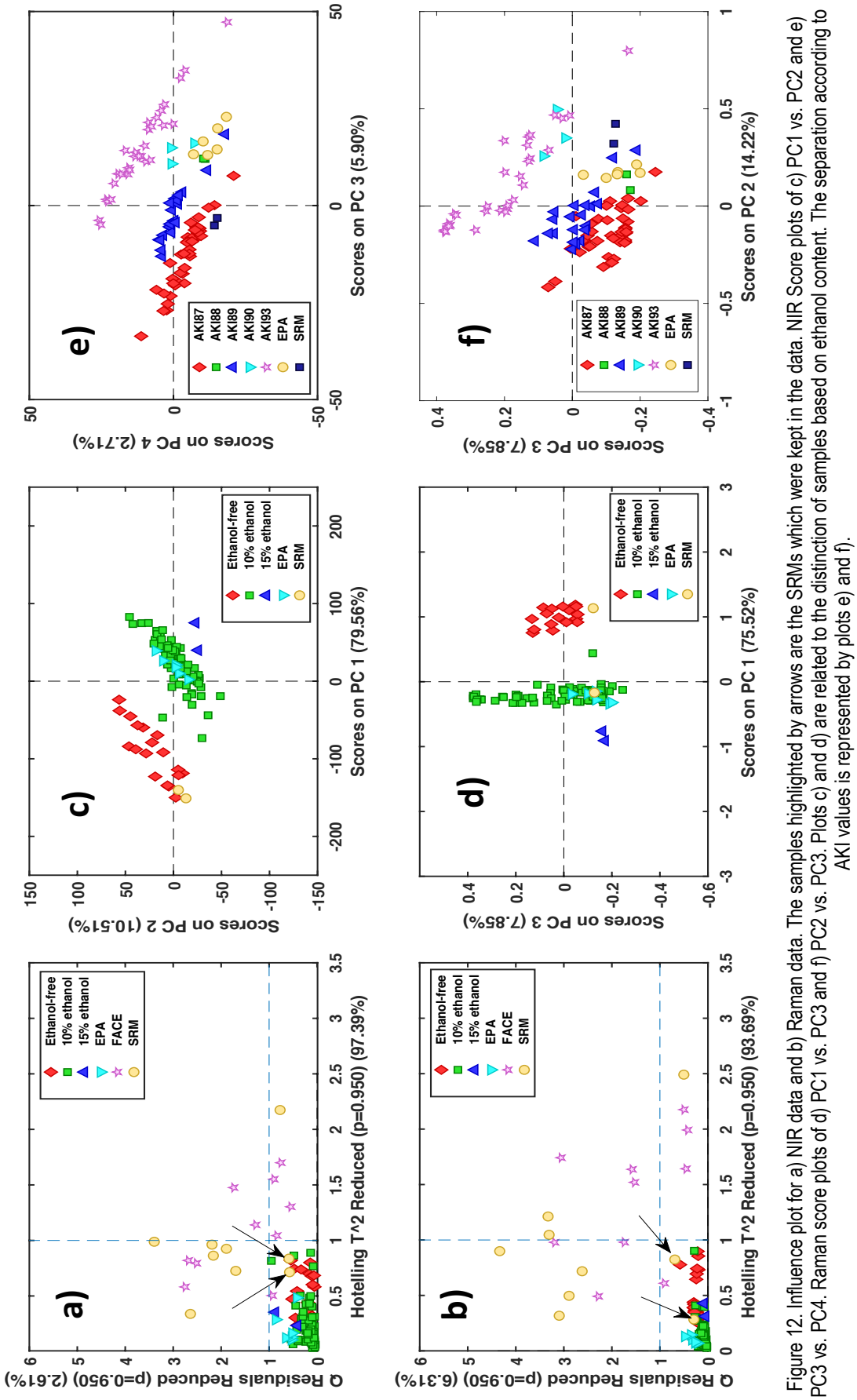

Figure 12. Influence plot for a) NIR data and b) Raman data. The samples highlighted by arrows are the SRMs which were kept in the data. NIR Score plots of c) PC1 vs. PC2 and e) PC3 vs. PC4. Raman score plots of d) PC1 vs. PC3 and f) PC2 vs. PC3. Plots c) and d) are related to the distinction of samples based on ethanol content. The separation according to AKI values is represented by plots e) and f).

5.1.4. Soft Independent Modelling of Class Analogy

In order to obtain classification models by SIMCA, it is necessary to build an individual PCA model for each class, these individual PCA models are constructed independently from each other, meaning that the number of PCs and type of preprocessing are determined individually for each class focusing on separating the class being analyzed from the others.

In order to perform the classification of gasoline samples according to their ethanol content and AKI values by SIMCA, the samples were split into calibration (CAL) and validation (VAL) sets, with the validation set consisting of 33% of the samples for each of the classes. The selection of samples for the validation set was carried out based on the distribution of PCA scores presented in the previous section, it was performed trying to select samples around the whole cluster for that specific class. The samples from classes which contain less than 4 samples, EPA samples, and SRMs were included in the validation set. Table 2 shows the splitting of samples based on specified classes. The SRMs were not included in the classification by AKI because their AKI values are not known.

Table 2. Splitting of gasoline samples into CAL and VAL sets for the classification by SIMCA.

		Ethanol Content				
-		**Ethanol-free**	**10% ethanol**	**15% ethanol**	**EPA (10% eth)**	**SRM**
CAL (69 samples)		13 samples	56 samples	-	-	-
VAL (38 samples)		5 samples	23 samples	2 samples	6 samples	2 samples
		Antiknock Index - AKI				
-	**AKI 87**	**AKI 88**	**AKI 89**	**AKI 90**	**AKI 93**	**EPA (AKI 88)**
CAL (69 samples)	32 samples	-	16 samples	-	21 samples	-
VAL (34 samples)	10 samples	2 samples	5 samples	3 samples	8 samples	6 samples

An important approach to obtain a good SIMCA classification model consists of evaluating the influence plot of each PCA model. The influence plot shows the separation between the samples of the class being modelled and the samples from the remaining classes based on *Q* residuals and Hotelling's T^2. The greater the separation observed by the influence plots the better the PCA model to be used by SIMCA. It is very important to use an adequate number of PCs in the PCA models in order to avoid overfitting of the classification model. Table 3 shows the information about the PCA models used to build SIMCA classification models.

Table 4 and Figure 13, and Table 5 and figure 14 show the results obtained by SIMCA models for ethanol content and AKI, respectively. The classification errors were limited to samples which were classified incorrectly classified as unassigned. Among the VAL sample sets there are samples which were correctly classified as unassigned including samples with *15% ethanol* from the ethanol content classification models and classes *AKI 88, AKI 90* and *EPA* from the AKI classification models. These samples were purposely included in VAL sets in order to demonstrate the capability of SIMCA to not misclassify samples that do not belong to classes for which models were constructed. The SRMs analyzed include SRM_2287 which contains 10% ethanol and SRM_2287b, which is a gasoline blank that is ethanol-free. The SIMCA models using NIR were not able to classify these SRMs satisfactorily while the Raman based model classified SRM_2287 correctly. This difficulty in classifying the SRMs is likely related to the point previously discussed in

the outlier diagnostics related with the compositional differences between SRMs and commercial gasoline samples. Regarding the samples that were incorrectly unassigned, most of them coincide by both NIR and Raman as it is demonstrated in figures 13 and 14 by labeled samples. The explanation about samples GAS102 and GAS113 was found, a mistake was noted on the information provided about sample GAS102 and this sample actually contains 5% ethanol thus being correctly classified as unassigned. Sample GAS113 spent more time at ambient temperature before being refrigerated than most of the remaining samples and it is possible that the composition of the fuel was compromised. The EPA samples contain 10% ethanol with AKI equal to 88, so, they were correctly classified by both classification analyses and spectroscopic techniques.

Table 3. Information about the PCA models used to build SIMCA classification models.

NIR DATA				
-	**Class**	**Preprocess**	**Number of PCs**	**Cumulative Variance**
Antiknock Index - AKI	AKI 87	SNV+MC	3	99.88%
	AKI 89	SNV+MC	3	99.31%
	AKI 93	1st der+MC	4	99.51%
Ethanol content	Ethanol-free	MC	2	96.73%
	10% ethanol	MSC+MC	1	72.54%
RAMAN DATA				
Antiknock Index - AKI	AKI 87	SNV+MC	4	96.57%
	AKI 89	SNV+MC	3	95.41%
	AKI 93	SNV+MC	3	97.94%
Ethanol content	Ethanol-free	MC	2	94.84%
	10% ethanol	MC	3	95.68%

The figures of merit were obtained after removing samples GAS102 and GAS113 from the results. Table 6 shows the figures of merit calculated for the SIMCA models to classify samples based on ethanol content and AKI values using both NIR and Raman spectroscopies. The general misclassification error (GME) was calculated for each set, which consists of the weighted average of the ME values. In general, the results obtained for the classification were very satisfactory using either spectroscopy method with both techniques having general classification errors less than 5.0% with sensitivity, precision, and specificity values equal to or very close to 1. Those results show to be very acceptable when compared to results for gasoline classification found on the literature. Ardila et al. (2017) classified gasoline samples according to their origin using Raman and PLS-DA with cross-validation errors less than 5.0%. De Oliveira et al. (2004) [78] used data from gasoline distillation curves was used with SIMCA to detect adulteration in Brazilian gasolines with 92% correct classification rate. Balabin et al. (2010) obtained average errors of 13 ± 1%, 4 ± 3%, and 11 ± 2% using SIMCA in a study comparing multivariate techniques for classifying gasoline samples according to their source and type by NIR.

Table 4. Confusion matrices for the classification of gasolines based on ethanol content by NIR and Raman.

-		Actual Class						
		Calibration		Validation				
		E00	E10	E00	E10	E15	SRM	EPA
NIR	Predicted as E00	13	0	5	0	0	0	0
	Predicted as E10	0	54	0	22	0	0	6
	Unassigned	0	2	0	1	2	2	0
	ME %	0.0	3.57	0.0	4.35	0.0	100.0	0.0
Raman	Predicted as E00	13	0	5	0	0	0	0
	Predicted as E10	0	55	0	21	0	1	6
	Unassigned	0	1	0	2	2	1	0
	ME %	0.0	1.79	0.0	8.70	0.0	50.0	0.0

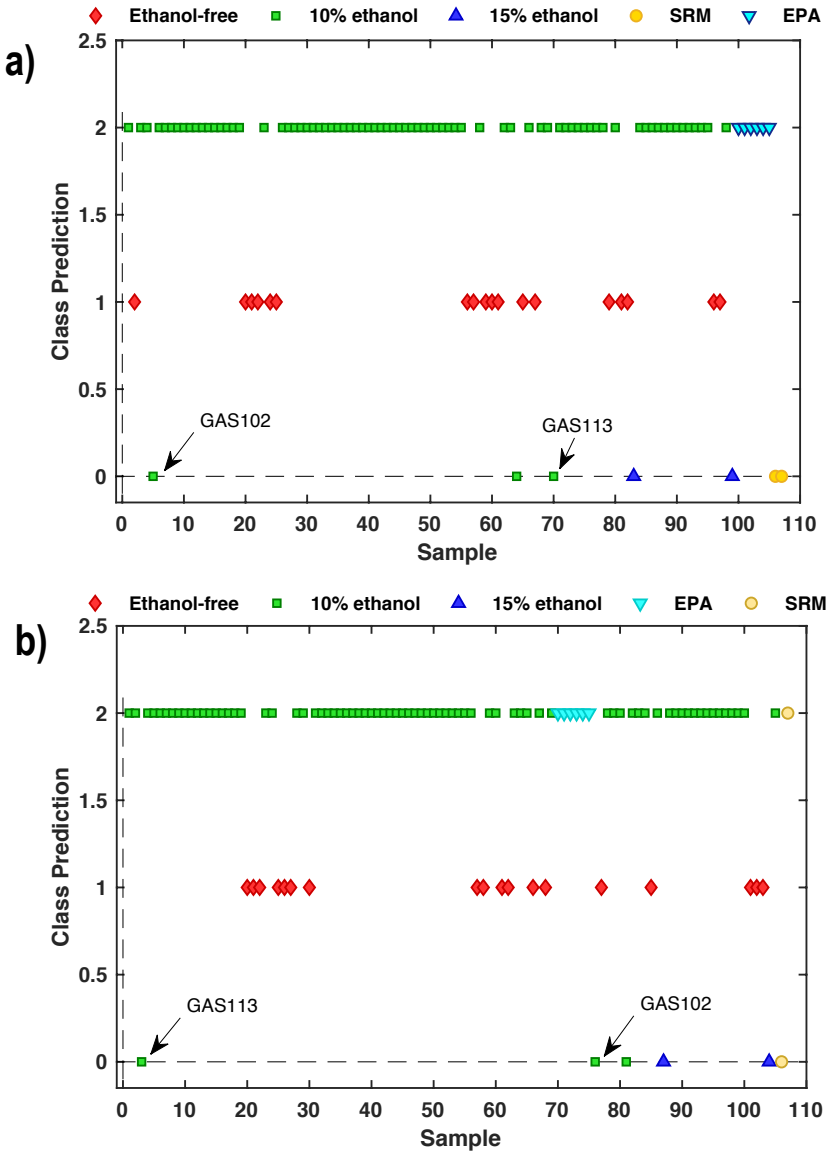

Figure 13. SIMCA Classification based on ethanol content of gasoline samples by a) NIR and b) Raman.

Table 5. Confusion matrices for the classification of gasolines based on AKI by NIR and Raman.

	-	Actual Class								
		Calibration			Validation					
		AKI 87	AKI 89	AKI 93	AKI 87	AKI 88	AKI 89	AKI 90	AKI 93	EPA
NIR	Predicted as AKI 87	33	0	0	10	0	0	0	0	0
	Predicted as AKI 89	0	16	0	0	0	5	0	0	0
	Predicted as AKI 93	0	0	22	0	0	0	0	7	0
	Unassigned	0	0	0	0	2	0	3	1	6
	ME %	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.5	0.0
Raman	Predicted as AKI 87	33	0	0	10	0	0	0	0	0
	Predicted as AKI 89	0	16	0	0	0	5	0	0	0
	Predicted as AKI 93	0	0	21	0	0	0	0	8	0
	Unassigned	0	0	1	0	2	0	3	0	6
	ME %	0.0	0.0	4.55	0.0	0.0	0.0	0.0	0.0	0.0

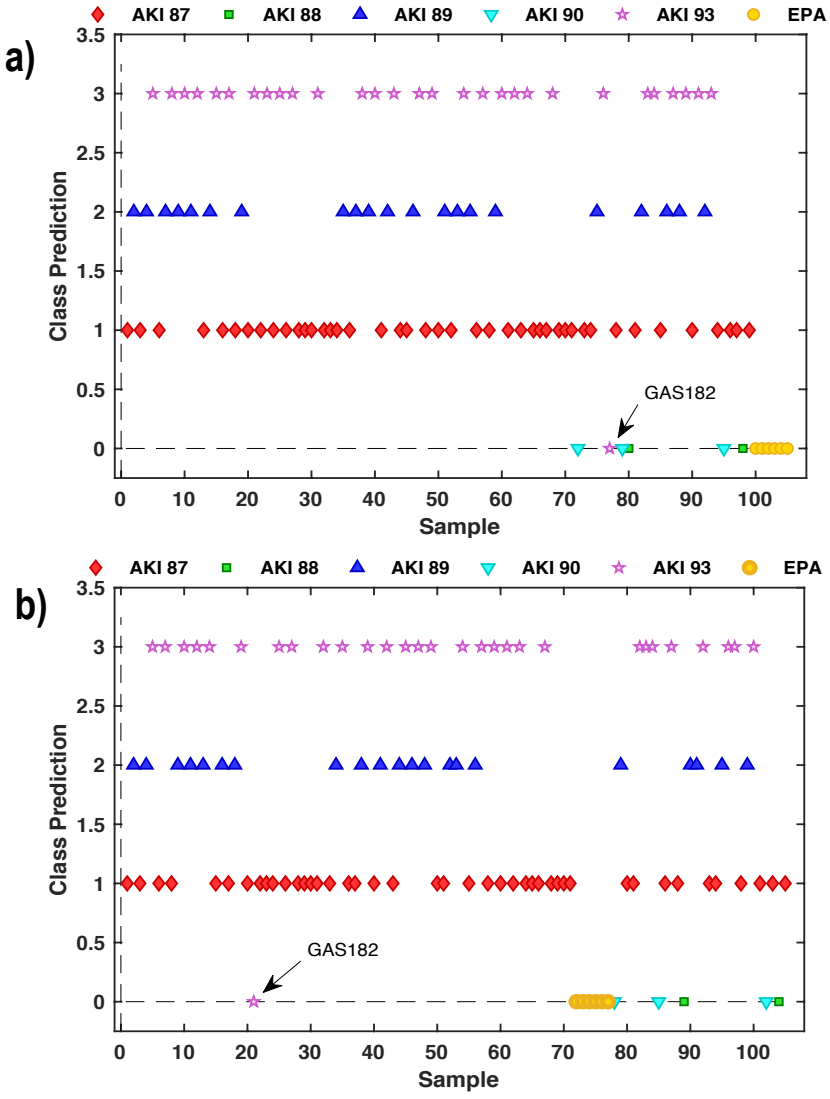

Figure 14. SIMCA Classification based on AKI of gasoline samples by a) NIR and b) Raman

Table 6. Figures of merit for SIMCA classification models

Ethanol content

-	NIR				Raman			
	Calibration		Validation		Calibration		Validation	
Class	**E00**	**E10**	**E00**	**E10**	**E00**	**E10**	**E00**	**E10**
ME %	*0.00*	*1.80*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*4.50*
Sensitivity	*1.00*	*0.98*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*0.95*
Precision	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
Specificity	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
GME %	*2.85%*				*1.89%*			

Antiknock Index-AKI

-	NIR						Raman					
	Calibration			Validation			Calibration			Validation		
Class	**AKI 87**	**AKI 89**	**AKI 93**	**AKI 87**	**AKI 89**	**AKI 93**	**AKI 87**	**AKI 89**	**AKI 93**	**AKI 87**	**AKI 89**	**AKI 93**
ME %	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*12.50*	*0.00*	*0.00*	*4.55*	*0.00*	*0.00*	*0.00*
Sensitivity	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*0.88*	*1.00*	*1.00*	*0.95*	*1.00*	*1.00*	*1.00*
Precision	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
Specificity	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
GME %	*0.95%*						*0.95%*					

5.2. Quantitative analysis

5.2.1. NIR bands assignment

For the quantitative analysis, NIR spectra were obtained in transmission mode using 5 mm pathlength quartz cuvettes. Figure 15a is a plot of the NIR transmission spectra of all 166 samples used for the quantitative analysis of gasoline. The spectra consist of the following bands: band related to the second overtone of CH bonds in the region between 1120-1250 nm, combination of CH_3 and CH_2 bands between 1350-1500 nm, the region comprised of first overtones of OH bonds is observed between 1500-1595 nm, first overtones of CH, CH_2, and CH_3 compose the band located between 1600-1900 nm, the combination of C=C is seen around 2000-2200 nm [75]. A significant difference in the transmission data set compared to the NIR transflectance spectra obtained for the qualitative analysis is the additional information available in the region between 1670-1790 nm, which was saturated in the transflectance data. This was simply a result of the shorter effective optical pathlength with the 5mm cuvette.

5.2.2. Raman bands assignment

Figure 15c shows the Raman spectra obtained for all 166 samples used for the quantitative analysis. These spectra consist of 830 nm Raman spectra covering the Raman shift range from 300 to 1600 cm^{-1}. This region often referred to as the fingerprint region, is composed largely of C-C skeletal vibrations and C-H deformations. The region between 250-400 cm^{-1} is related to C-C bending of aliphatic; C-C stretching of aliphatic is located between 600-1300 cm^{-1}, where, C-C stretching for aromatics is found around 1000 cm^{-1}; the region comprised of CH_2 and CH_3 bending is observed between 1400-1470 cm^{-1} [76].

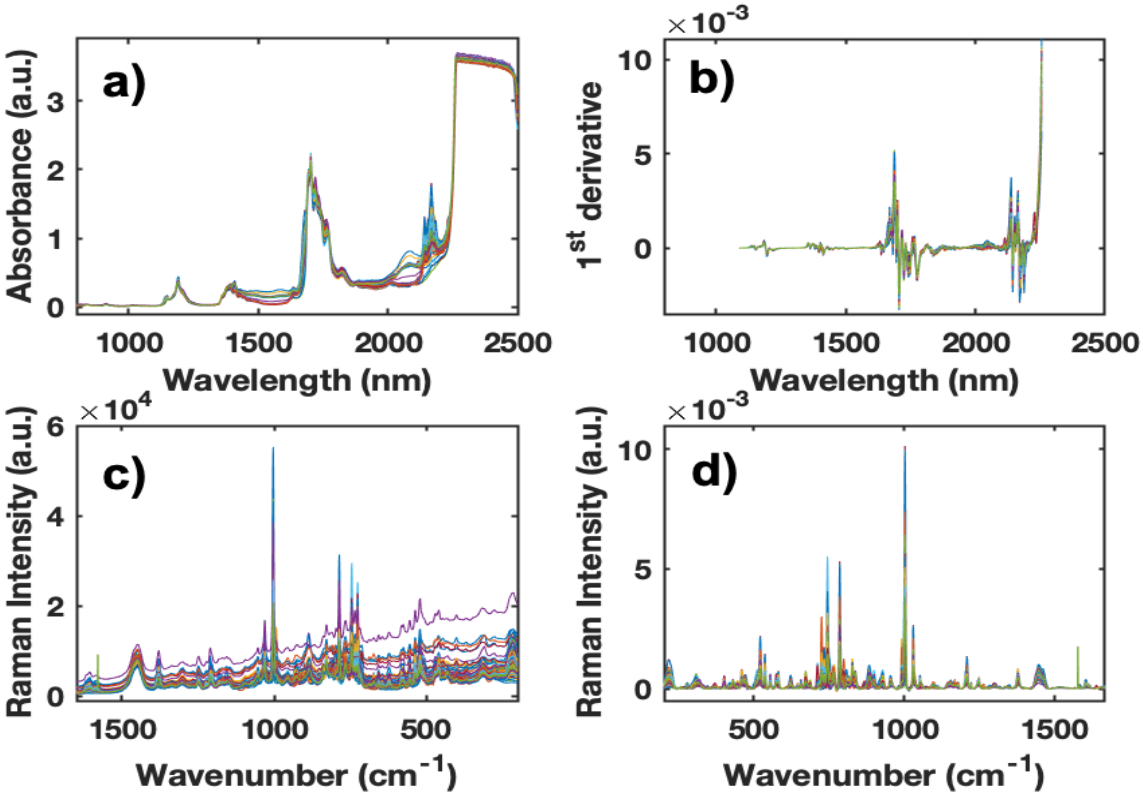

Figure 15. Original spectra obtained by a) NIR and c) Raman. Preprocessed spectra for b) NIR and d) Raman.

5.2.3. Quantitative ^{1}H NMR reference method

The sample set used for the quantitative analysis consists of NAFS, EPA, FACE, and commercial gasoline samples (GAS). However, reference values for gasoline composition were not available for GAS samples. In addition, the only hydrocarbon class data available for the NAFS samples was from the ASTM D-1319 FIA method, which is known to have issues with interferences from oxygenates and large uncertainties based on reproducibility. To provide quantitative hydrocarbon class it was necessary to carry out a quantitative analysis of all samples to obtain reference values. A method based on proton nuclear magnetic resonance (^{1}H NMR) described in the literature [73,74] was chosen to be used as reference method for that purpose. This ^{1}H NMR method provides quantitative gasoline composition values for aromatics, olefins, paraffins, benzene, ethanol, and MTBE.

The ^{1}H NMR spectra for all 166 samples used for quantitative analysis were acquired followed by Fourier transformation, phase correction, baseline correction, and chemical shift referencing to TMS. The quantitative composition results were obtained according to the equations previously presented in section *4.2.2.a*. A wide range of compositional variation was observed for aromatics, olefins, paraffins, and benzene, which provided a good basis for the development of quantitative calibration models. A predictive model for ethanol content would be of interest; however, the available samples were almost universally in two categories, ethanol-free and 10% ethanol. This distribution was considered unsuitable to develop and test models for predicting ethanol content and therefore this was not investigated. A paired *t*-test was performed in order to compare the results of gasoline composition obtained by ^{1}H NMR and FIA (ASTM D-1319) for the NAFS samples. The results are shown in Table 7. While the results from the two methods are similar, the hypotheses for all paired *t*-tests were rejected, indicating that these two methods are not commutable. It does not necessary mean that ^{1}H NMR is not a good method, Beens et al. (2003) showed some drawbacks that can occur in analysis performed by ASTM D-1319 which can generate erroneous results with a wide bias. Some samples were analyzed in replicates in order to assess the precision of the ^{1}H NMR method. These results are presented in Table 8, which demonstrates that the precision of the ^{1}H NMR was quite good, and the method was considered reliable for the purposes of this work. Biases and uncertainties with the NMR reference method were not investigated further.

Table 7. Paired *t*-test for comparison of results obtained by ^{1}H NMR and ASTM D-1319.

Parameter	Number of samples	Average of difference	Standard deviation of difference	t-calculated	t-critical α=0.05	Hypothesis
Aromatics	*100*	*1.309*	*1.525*	*8.58*		*Rejected*
Olefins	*100*	*0.341*	*1.004*	*3.40*	*1.98*	*Rejected*
Paraffins	*100*	*1.662*	*1.990*	*8.35*		*Rejected*

5.2.4. Principal Component Analysis

PCA was carried out in order to perform outlier diagnostics on the NIR and Raman data sets. The NIR spectra were preprocessed using derivative Savitzky-Golay (1st derivative, 2nd order polynomial, 9 points window) followed by normalization (by area) and mean centering. The Raman spectra were preprocessed using a combination of baseline correction (Whittaker method , $p = 0.01, \lambda = 1000$), normalization (by area), and mean centering. The preprocessed spectra for NIR

and Raman are shown in figures 15b and 15d, respectively. Baseline and intensity fluctuation effects could be successfully corrected in both cases. PCA models were built with 4 PCs for NIR, accounting for 91.47% of the variance, and 3 PCs for Raman, accounting for 86.98% of the variance. The outlier diagnostics were performed by examining the influence plots which are shown in figures 16a and 16b for NIR and Raman, respectively. Some of the EPA, FACE, and NAFS (from Mexico, Canada, and Alaska) samples contained high residuals. This high residual can be related to the under-representation of similar samples to these in the data set. It was decided to keep these samples in both the NIR and Raman data sets and then evaluate them during the construction of PLS models.

Table 8. Precision of ^{1}H NMR results.

Sample ID	Number of replicates	Aromatics		Olefins		Paraffins	
		Average (% v/v)	RSD %	Average (% v/v)	RSD %	Average (% v/v)	RSD %
NA001	*3*	*4.914*	*0.272*	*1.980*	*0.340*	*93.028*	*0.018*
NA006	*3*	*12.034*	*0.298*	*3.697*	*0.557*	*75.066*	*0.025*
NA016	*3*	*28.229*	*0.041*	*5.071*	*0.258*	*57.202*	*0.018*
NA030	*4*	*16.237*	*0.727*	*6.864*	*0.573*	*67.683*	*0.088*
NA048	*3*	*19.226*	*0.063*	*3.484*	*0.199*	*68.324*	*0.025*
NA054	*4*	*27.121*	*0.049*	*9.367*	*0.229*	*53.739*	*0.045*
NA062	*3*	*22.901*	*0.053*	*10.622*	*0.136*	*56.834*	*0.042*
NA074	*3*	*24.285*	*0.081*	*1.884*	*1.131*	*64.251*	*0.011*
NA081	*4*	*20.195*	*0.604*	*8.194*	*0.440*	*62.302*	*0.103*
NA108	*4*	*17.924*	*0.133*	*11.162*	*0.227*	*61.324*	*0.050*

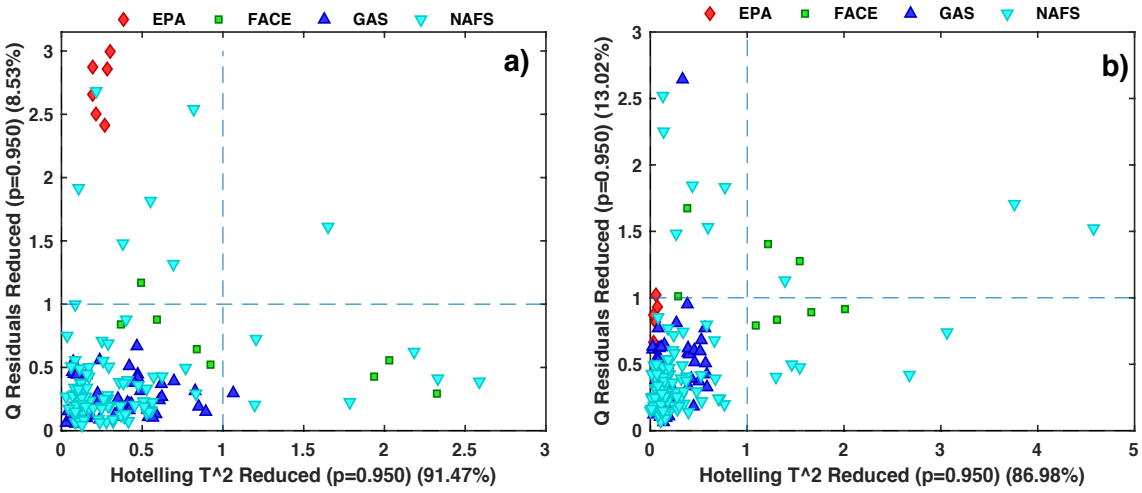

Figure 16. Influence plot for a) NIR and b) Raman data.

5.2.5. Prediction of gasoline parameters using Partial Least Squares

The first step in PLS model building for predicting gasoline properties consisted of determining the best preprocessing strategy. To do that, different preprocessing strategies were applied to the NIR and Raman data sets, and PLS models using all available samples for each parameter were built. Those PLS models were then compared by observing the cross-validation results, RMSECV. For NIR, the lowest RMSECVs were obtained with a combination of 1st derivative (Savitzky-Golay method, 2nd order polynomial function, 9 points window), normalization (by area)

and mean centering were chosen to preprocess NIR data. For the Raman data set, the best RMSECVs were obtained with a combination of baseline correction (Whittaker method, $p = 0.01$, $\lambda = 1000$), normalization (by area) and mean centering.

After optimizing preprocessing, the samples were split into two sets: a calibration (CAL) set used to construct the PLS calibration models and a validation (VAL) set consisting of external samples used to evaluate the performance of PLS models. The splitting was performed such that ~30% of the samples were selected for the VAL set. The VAL set samples were carefully chosen in order to obtain the maximum representativity of the parameter ranges and avoiding the selection of extreme samples.

PLS models were built for the prediction of 8 gasoline parameters. The number of latent variables was selected according to the lowest value of RMSECV found before additional latent variables either provide no improvement or are detrimental to the performance as exemplified in figure 17. Table 9 summarizes the PLS model results obtained for prediction of relative density, MON, RON, AKI, Aromatics, Benzene, Olefins, and Paraffins by both NIR and Raman spectroscopies.

Table 9. PLS models results for 8 gasoline parameters by NIR and Raman spectroscopies.

Gasoline Property	Technique	Units	LVs	R^2_C	RMSEC	ARE_C (%)	R^2_{CV}	RMSECV	R^2_P	RMSEP	ARE_P (%)
Relative Density	**NIR**	**g/cm³**	*9*	*0.976*	*0.0014*	*0.15*	*0.914*	*0.0027*	*0.979*	*0.0015*	*0.17*
	Raman	**g/cm³**	*8*	*0.950*	*0.0020*	*0.21*	*0.883*	*0.0032*	*0.948*	*0.0022*	*0.23*
MON	**NIR**	**-**	*7*	*0.981*	*0.3455*	*0.32*	*0.970*	*0.4360*	*0.984*	*0.3127*	*0.31*
	Raman	**-**	*7*	*0.948*	*0.5642*	*0.51*	*0.900*	*0.7942*	*0.971*	*0.4517*	*0.41*
RON	**NIR**	**-**	*10*	*0.983*	*0.3683*	*0.31*	*0.961*	*0.5571*	*0.978*	*0.3838*	*0.33*
	Raman	**-**	*10*	*0.961*	*0.5255*	*0.45*	*0.892*	*0.8321*	*0.941*	*0.6739*	*0.57*
AKI	**NIR**	**-**	*9*	*0.985*	*0.3069*	*0.28*	*0.969*	*0.4338*	*0.995*	*0.1970*	*0.17*
	Raman	**-**	*10*	*0.974*	*0.4184*	*0.37*	*0.923*	*0.7117*	*0.983*	*0.3367*	*0.32*
Aromatics	**NIR**	**% v/v**	*10*	*0.997*	*0.4310*	*2.11*	*0.995*	*0.5850*	*0.998*	*0.3862*	*1.65*
	Raman	**% v/v**	*6*	*0.994*	*0.6512*	*2.65*	*0.988*	*0.8981*	*0.994*	*0.6682*	*3.25*
Benzene	**NIR**	**% v/v**	*9*	*0.992*	*0.0229*	*6.59*	*0.984*	*0.0315*	*0.996*	*0.0158*	*3.30*
	Raman	**% v/v**	*8*	*0.979*	*0.0350*	*5.23*	*0.963*	*0.0459*	*0.983*	*0.0360*	*1.19*
Olefins	**NIR**	**% v/v**	*8*	*0.987*	*0.3316*	*7.47*	*0.979*	*0.4246*	*0.987*	*0.3545*	*10.98*
	Raman	**% v/v**	*12*	*0.981*	*0.4292*	*8.94*	*0.929*	*0.8201*	*0.976*	*0.4436*	*6.56*
Paraffins	**NIR**	**% v/v**	*8*	*0.996*	*0.5587*	*0.76*	*0.993*	*0.7286*	*0.998*	*0.3505*	*0.47*
	Raman	**% v/v**	*7*	*0.956*	*1.6932*	*2.07*	*0.937*	*2.1219*	*0.968*	*1.6097*	*2.16*

a) Relative density

The NAFS samples were used to build PLS models for determining relative density, however, two of these samples exhibited extreme reference values when compared to the remaining samples, so they were removed from the data set. A total of 98 NAFS samples were used with 67 samples included in the CAL set and 31 samples in the VAL set.

Both spectroscopic techniques presented very similar errors with NIR performing slightly better than Raman. The response residuals are plotted in figures 19a and 19b and no outliers or trends were observed with either technique. Alternative methods for determining relative density of

gasoline are also found in the literature. Flumignam et al. (2008) [79] have obtained values of 0.0067 g/cm^3 for RMSEC and 0.0043 g/cm^3 for RMSEP using gas chromatography (GC) profiles and PLS models resulting in an ARE of 0.55%. Godoy et al. (2011) [80] obtained a RMSEP of 1.7 Kg/m^3 (0.0017 g/cm^3) using comprehensive two-dimensional gas chromatography and PLS. The results obtained in this work by both NIR and Raman demonstrate to be very good when compared with the ones reported in the literature, it can be confirmed by observing correlation and EJCR plots in figures 21a-b and 23a, respectively.

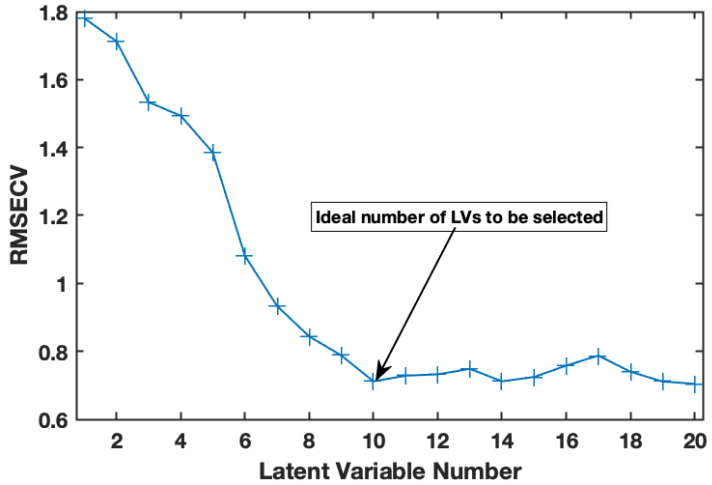

Figure 17. Demonstration on how the optimum number of LVs is determined for a PLS model according to the lowest value of RMSECV just before its variation starts to be almost constant.

b) Octane parameters: MON, RON, and AKI

The PLS models for determining octane ratings (MON, RON, AKI) of gasoline were calculated using the 100 NAFS samples with 67 samples in the CAL set and 33 samples in the VAL set. The best NIR results were obtained by selecting the spectral region between 1100-1300 nm for the PLS models. This region corresponds to the second overtone of CH bonds of aromatics and methylene which are related to the octane numbers [42].

Some outliers were detected for those parameters by both techniques as it can be seen in figures 19c-h by labeled samples. However, these outliers coincide for octane parameters by both techniques, it indicates that some error has occurred during their analyses by the reference method, as it was not possible to reanalyze these samples by the reference method, they have been removed. Furthermore, no trends were observed in the response residuals plots.

While both NIR and Raman methods produced good results, error were lower for NIR. Both results were compared with the ones reported in the literature. Swarin et al. (1991) [81] have used NIR and PLS for predicting 16 gasoline properties, SEP results of 0.370 and 0.353 were obtained for MON and RON, respectively. Cooper et al. (1995) [44] obtained SEV values of 0.415, 0.535, and 0.410 for MON, RON, and AKI, respectively, by using FT-Raman spectroscopy and PLS regression. In a more recent study carried out by Voigt et al. (2019), RMSEC of 0.23 and RMSEP of 0.78 were obtained for the prediction of RON by using PLS and a handheld Raman spectrometer. Results for determining MON and RON by different instruments are also found in the literature [11,79,82]. The results obtained in this work were similar or somewhat better than results reported in the literature with one exception being for the results obtained by Mendes et al. (2012) that

presented RMSEP values of 0.085 for RON and 0.063 for MON. Figures 21c-h and 23b-d demonstrate that the use of either NIR or Raman spectroscopies with PLS regression present very reliable results for determining octane parameters.

Furthermore, in order to demonstrate the reliability of MON and RON results predicted by the PLS models, those predicted values were used to calculate the AKI values, and the results were compared with the reference values. As shown in figures 18a-c, a very good correlation is presented when comparing AKI reference values versus AKI values obtained by MON and RON predicted by PLS models. In addition, the EJCR ellipses contain the ideal theoretical point indicating that constant and proportional bias are absent.

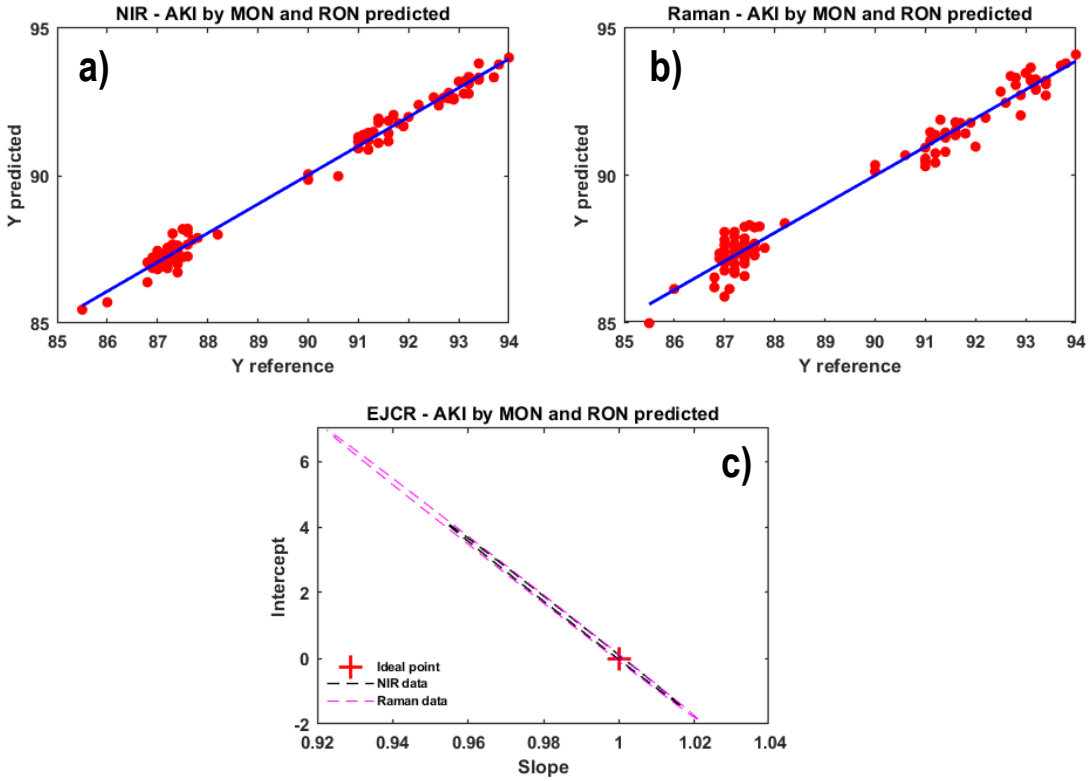

Figure 18. Correlation plots of reference values of AKI vs. AKI values obtained using MON and RON values predicted by PLS models using a) NIR and b) Raman data. c) EJCR for reference values vs. predicted values. If the ideal theoretical point (0,1) lies inside the confidence ellipse, BIAS is absent.

c) Aromatics

NIR and Raman spectra from all samples present in the quantitative data set were used to construct PLS models to predict gasoline aromatic composition. The CAL set was composed of 116 samples while the VAL set consisted of 50 samples. The selection of specific spectral regions was not needed to obtain the best results.

The response residuals can be observed in figures 20a-b, only one outlier was detected for the model constructed using Raman data, which corresponds to a FACE sample with ethanol 10%. This sample is under-represented in the data set and the PLS model was not able to fit it well using the Raman spectroscopy. This sample was discarded from the data. Despite this, the response residuals from both techniques were normally distributed and no trend were observed.

The results obtained for the prediction of gasoline composition by aromatics were excellent with RMSEP values less than 0.70 %v/v. It can also be confirmed by observing the correlation and EJCR plots in figures 22a-b and 23e. Aromatics content was also determined by Swarin et al. (1991) by using NIR and PLS resulting in a SEP value of 1.71 %v/v. Flumignan et al. (2008) got a RMSEC of 1.7 %v/v and RMSEC of 2.2 %v/v by PLS models developed with GC data. The results for the errors obtained in this work were less than the ones reported on the literature, it indicates that although errors presented by Raman results are higher than NIR's, both techniques are trustable to be used for predicting gasoline composition by aromatics.

d) Benzene

A total of 166 samples were available for predicting benzene content in gasoline. Two samples exhibited much higher benzene content when compared to the remaining samples and they were removed from the data set. The samples were divided into CAL and VAL sets containing 114 and 50 samples, respectively. The selection of specific spectral regions was performed in both NIR and Raman spectra to obtain the best results. For the NIR data, the region corresponding to the first overtones of CH bonds of aromatics which is located between 1650-1850 nm was selected while the region between 700-1120 cm^{-1} which contains C-C stretching for aromatics around 1000 cm^{-1} was selected for the Raman data [83].

Figures 20c-d show the response residuals obtained by the PLS models. No trends are observed in these plots, but one outlier was detected for each spectroscopic technique. These outliers are possibly related to some spectral issue which could have affected with more intensity the spectral regions selected to predict Benzene. These samples were eliminated from the data sets.

The performance for predicting benzene by NIR and Raman are similar with Raman presenting results modestly better than the ones obtained by NIR. It should be noted that the ARE values were quite high if compared with the ones obtained by PLS models of other parameters (except olefins). This is because the benzene content by volume is typically quite low, <1.3%, which results in larger relative errors compared to the bulk hydrocarbon class content. Nevertheless, it can be seen in figures 22c-d and 23f that the results obtained by both techniques are very good even if compared with the results reported on the literature. Flumignan et al. (2008) reported a RMSEC value of 0.0736 %v/v and RMSEP of 0.0919 %v/v resulting in an ARE of 5.21% using GC and PLS. Cooper et al. (1997) [41] used PLS models to determine BTEX (benzene, toluene, ethylbenzene and xylene) with different vibrational techniques and reported SEP values of 0.101 %w/w, 0.190 %w/w, and 0.088 %w/w by Raman, NIR, and Mid-IR, respectively. Thus, the results obtained in this work either by NIR or Raman are comparable with other studies reported in the literature.

e) Olefins

For the olefin analysis, one extreme sample with an unusually high olefin content was removed from the data set and a total of 165 samples were used to build the PLS models. Samples were sorted into CAL and VAL groups which contained 115 and 50 samples, respectively. In order to obtain the best results using Raman, the spectral region between 1125-1530 cm^{-1} was selected.

Figure 20e-f show a normal distribution for the response residuals obtained from the PLS models by NIR and Raman. One outlier can be observed in the response residual plot for the results obtained from NIR. This corresponds to an extreme sample with the second highest olefin concentration of the raw data set (166 samples). Because of this high concentration it was hard to fit this sample in the PLS model obtained using NIR spectroscopy. This sample was excluded from the NIR data set.

Olefins results present the highest values for ARE among all gasoline parameters being studied in this work, which is likely due to the lower levels of compounds in this class in gasoline samples with some containing less than 0.1%. This results in higher relative errors compared to the other hydrocarbon classes. In general, the results obtained by NIR and Raman were very similar with one exception for the values of ARE, but again, the prediction of samples with low olefin content has a big influence on these values. From figures 22e-f and 23g it can be seen that the results for olefin determination were very satisfactory by both techniques. Swarin et al. (1991) found a SEP error of 1.41 %v/v for the prediction of olefins using PLS and NIR. Flumignan et al. (2008) presented RMSEC and RMSEP of 4.3 %v/v and ARE of 25.60% obtained using GC data and PLS. Reboucas et al. (2011) [38] reported a RMSECV of 0.28 %w/w and two external validation sets were used resulting in RMSEP values of 0.42 %w/w and 3.10 %w/w using NIR and PLS. The results obtained in this work were better than the ones reported in the literature except for the results reported by Reboucas et al. (2011) that were similar.

f) Paraffins

All 166 samples were used to build the PLS models for determining paraffin content of gasoline. NIR and Raman spectra were split into CAL and VAL sets consisting of 116 and 50 samples, respectively. The spectral region of 860-1120 cm^{-1} was selected from the Raman spectra in order to obtain the best results. This spectral region is related to C-C skeletal stretching of paraffins.

The response residuals plots obtained for both techniques appear to be normally distributed and no trends were observed as evident in figures 20g-h. One outlier was found for the results obtained by NIR, which corresponds to one of the EPA samples. These were synthetically blended fuel samples and it is not surprising that an unreliable prediction could result. The PLS model using NIR data may not be robust enough to model this sample. This sample was removed from the data set.

Concerning the results obtained for paraffin quantification, the model obtained using Raman presented errors about twice as large as those obtained for the models using NIR data. However, the results obtained by both techniques are still satisfactory with values of ARE less than 2.5%. Figures 22g-h and 23h reveal the accuracy of the model with high correlation observed between the reference and predicted values is observed and both EJCR ellipses included the ideal theoretical point. A comparison with similar works reported in the literature was conducted. Swarin et al. (1991) reported a SEP error of 1.41 %v/v using PLS and NIR. Flumignan et al. (2008) obtained values of 5.0 %v/v and 4.5 %v/v for RMSEC and RMSEP, respectively, with an ARE of 8.50% using GC and PLS. Reboucas et al. (2011) using NIR and PLS obtained a RMSEC of 0.63 %w/w, and RMSEP values of 1.57 %w/w and 0.67 %w/w by using two external validation sets. The results obtained by NIR were better than the ones reported in the literature. While the results of the Raman

based model were better than the ones obtained by Flumignan et al. (2008) and similar to the results obtained by the others.

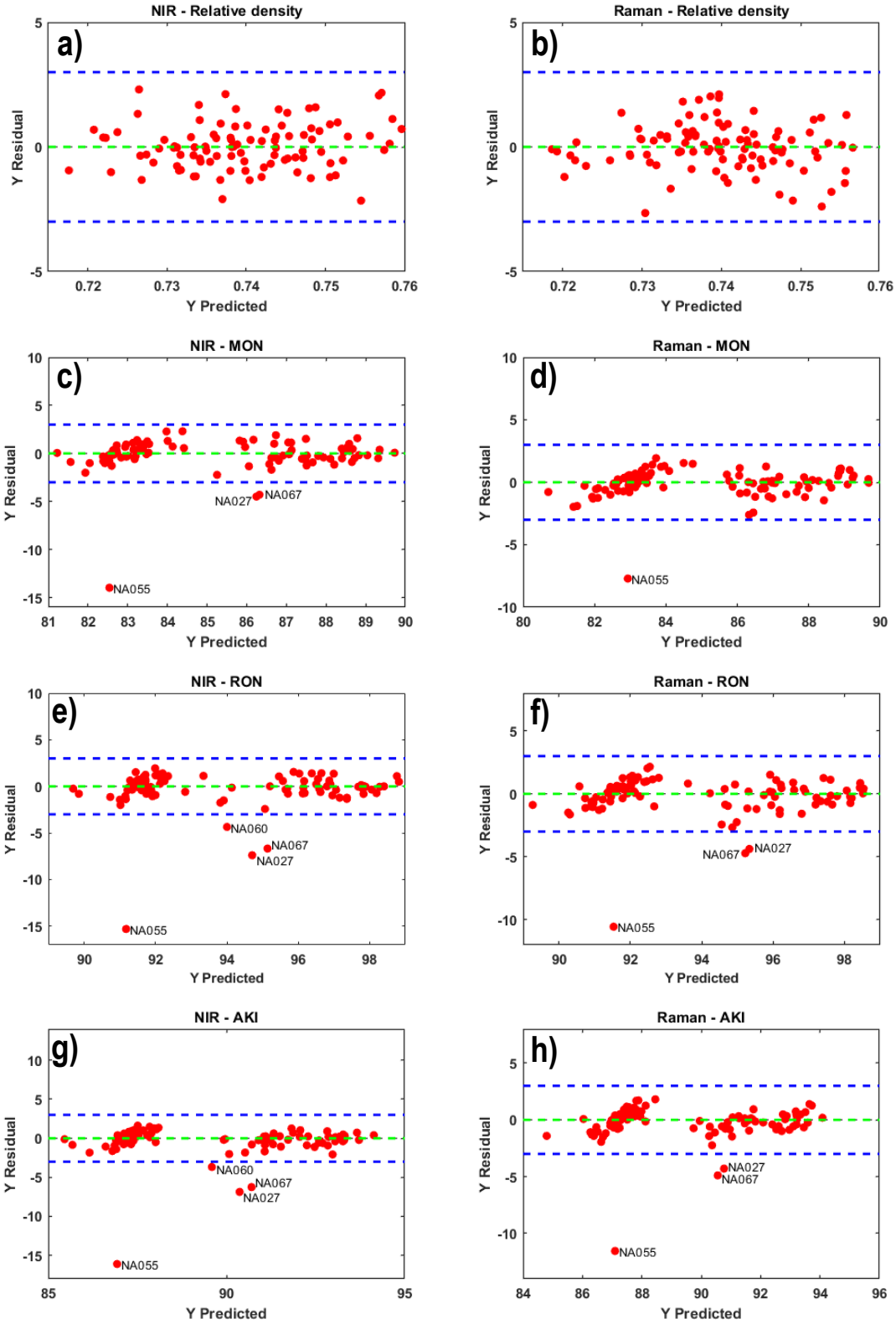

Figure 19. Response residuals for: Relative density by a) NIR and b) Raman; MON by c) NIR and d) Raman; RON by e) NIR and f) Raman; AKI by g) NIR and h) Raman. Labeled samples represent the outliers found for octane parameters which coincide by both NIR and Raman data.

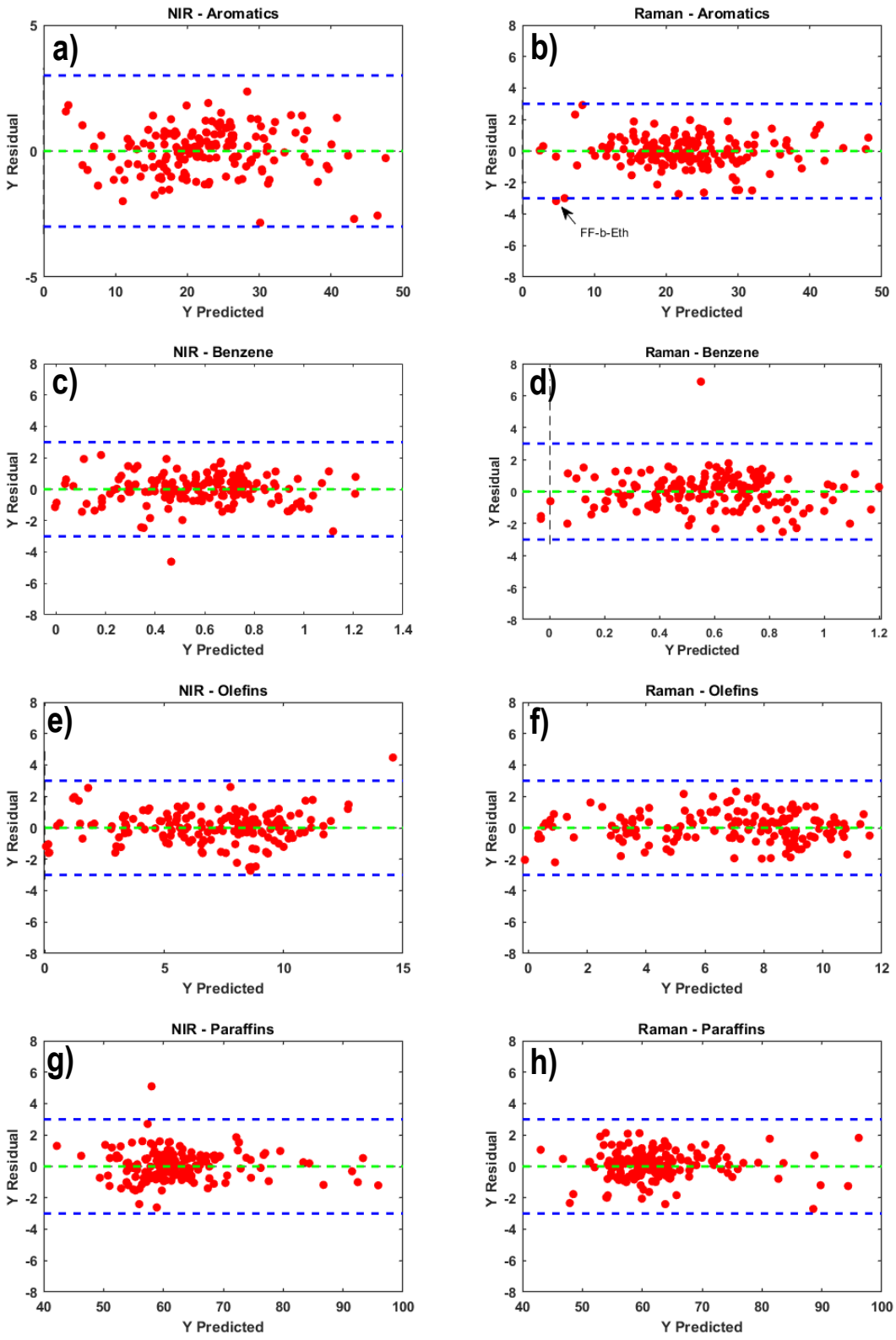

Figure 20. Response residuals for: Aromatics by a) NIR andb) Raman; Benzene by c) NIR and d) Raman; Olefins by e) NIR and f) Raman; Paraffins by g) NIR and h) Raman.

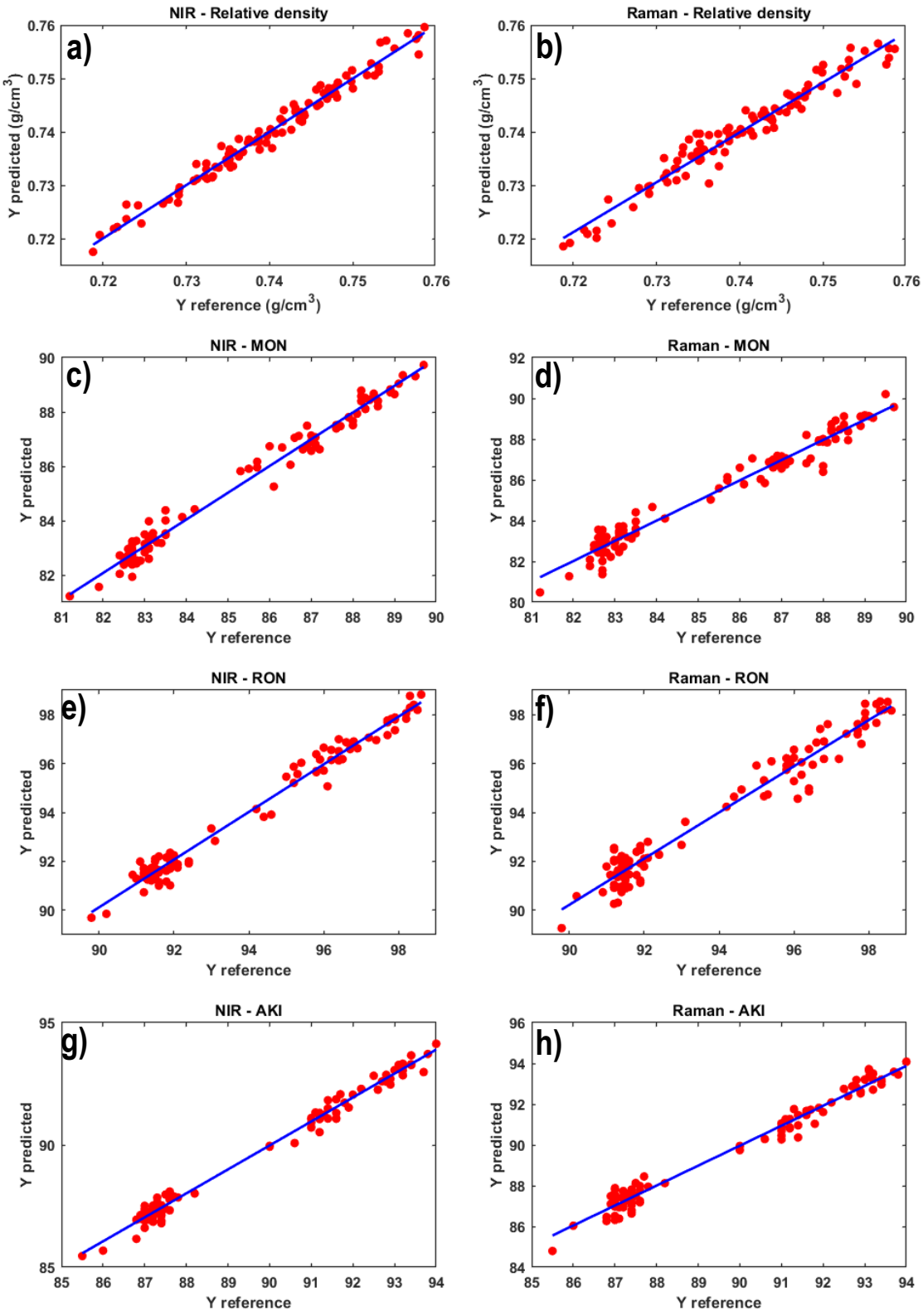

Figure 21. Correlation plots of reference vs. predicted values for: Relative density by a) NIR and b) Raman; MON by c) NIR and d) Raman; RON by e) NIR and f) Raman; AKI by g) NIR and h) Raman.

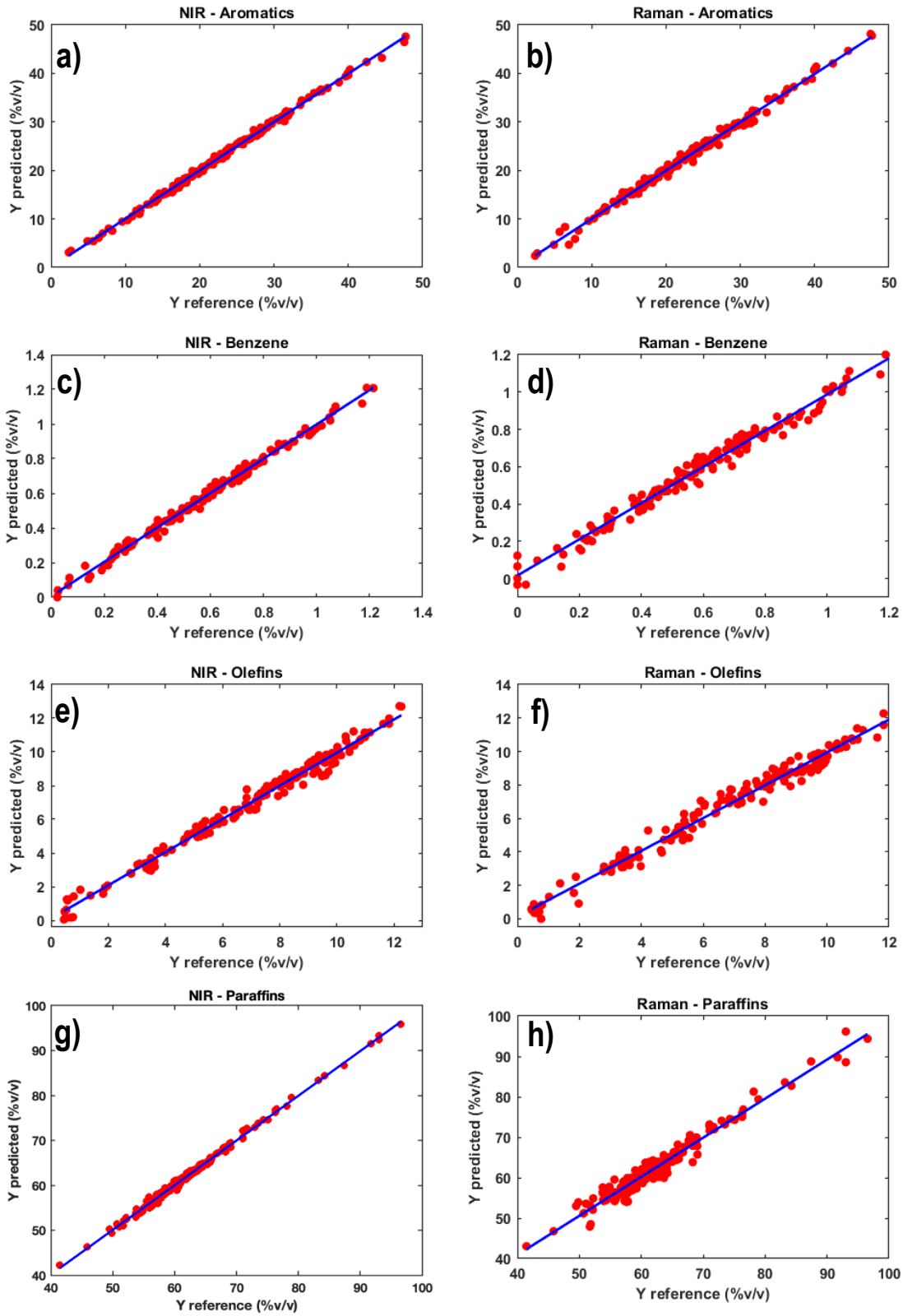

Figure 22. Correlation plots of reference vs. predicted values for: Aromatics by a) NIR andb) Raman; Benzene by c) NIR and d) Raman; Olefins by e) NIR and f) Raman; Paraffins by g) NIR and h) Raman.

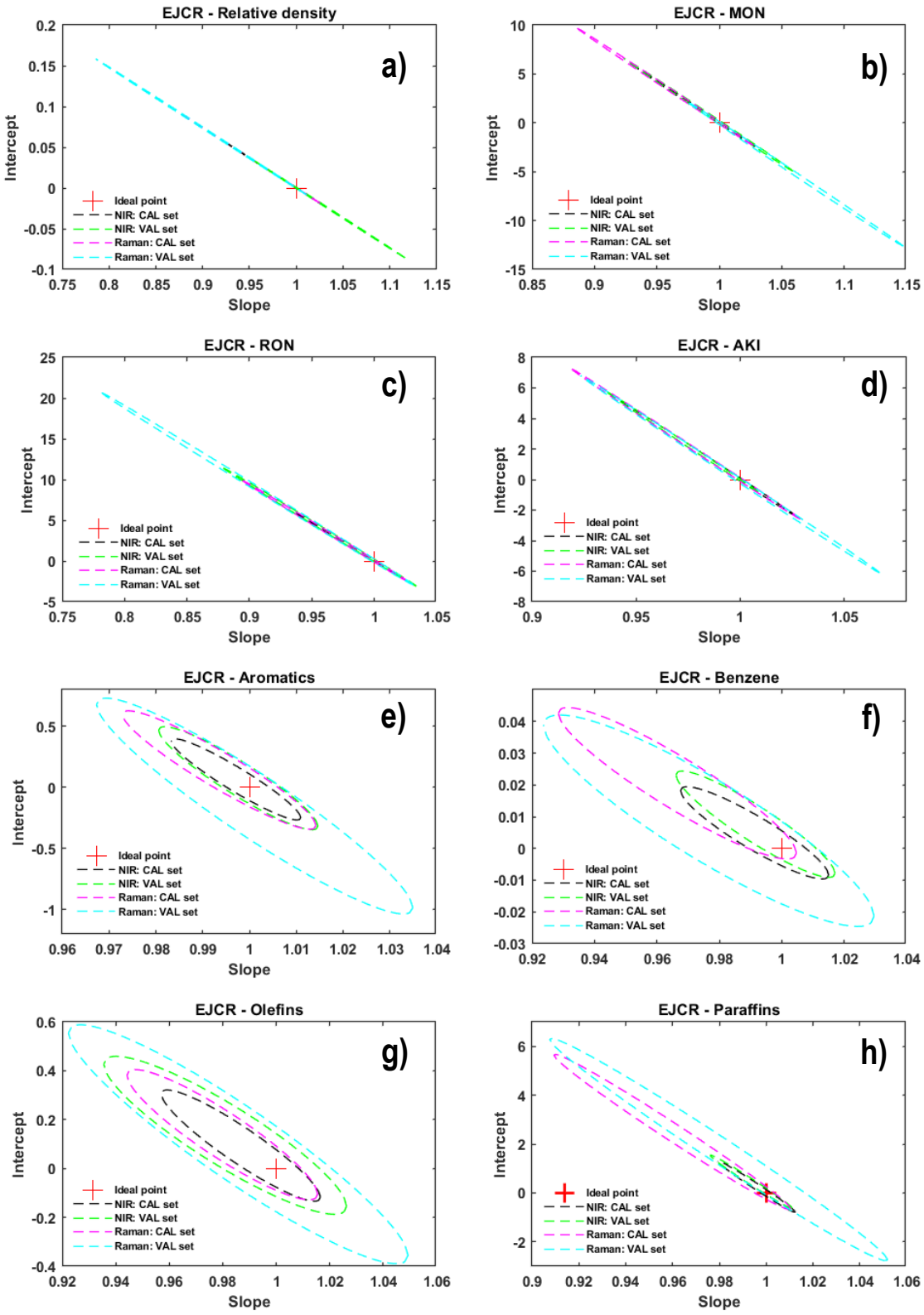

Figure 23.Elliptical joint confidence region for reference values vs. predicted values for: a) Relative density; b) MON; c) RON; d) AKI; e) Aromatics; f) Benzene; g) Olefins; h) Paraffins. If the ideal theoretical point (0,1) lies inside the confidence ellipse, BIAS is absent.

6. CONCLUSIONS

This study proposed to investigate the use of NIR and Raman spectroscopic techniques together with chemometric tools for developing alternative methods for the qualitative and quantitative analyses of motor gasoline samples based on chemical and physical properties.

Regarding qualitative analysis, SIMCA models were built for both NIR and Raman data in order to classify commercial gasoline samples based on their ethanol contents and AKI values. Good results were obtained in this work with only a few samples being incorrectly unassigned and no samples misclassification. Some of the samples incorrectly unassigned were verified and it was concluded that they represented outliers. After these samples were removed, prediction errors below 3% were obtained for both ethanol content and AKI for both techniques, which were better than results reported in the literature for similar studies. Comparing the SIMCA models obtained with NIR and Raman, the results were similar and both techniques would be suitable for classifying commercial motor gasoline samples based on ethanol content and AKI.

NIR and Raman data were also used to construct PLS models for the quantification of eight physical and chemical properties of gasoline, which include relative density, RON, MON, AKI, aromatics, benzene, olefins and paraffins. The performance of these PLS models was assessed based on RMSEC, RMSECV, RMSEP, and ARE values. The results obtained in this work were compared with other reports in the literature, and it was observed that for many parameters they were similar or better. EJCR tests were performed for evaluating the accuracy of the PLS models, the ellipses for all models including the ideal theoretical point (1,0) indicating that bias was absent. In general, the results obtained for the PLS models by both NIR and Raman spectroscopies were excellent with NIR data presenting better results than those for Raman.

The results obtained in this work show that the use of NIR and Raman spectroscopies with chemometric tools provide promising alternative methods for both qualitative and quantitative analyses of motor gasoline samples. These methods require little or no sample preparation, small amounts of sample, and are rapid. All these characteristics make them very attractive alternatives to the conventional methods. However, in order to use the methods developed in this work for the routine analysis of gasoline samples, it would be essential to expand the calibration of the SIMCA and PLS models by including new representative samples of different compositions, regions (cities, estates, countries), and seasons. This will increase the reliability of these models for the prediction of new unknown samples.

7. REFERENCES

[1] A. Szklo, V. Uller, M. Bonfá, Fundamentos do refino de petróleo: tecnologia e economia, 3.ed., Editora Interciência Ltda., Rio de Janeiro, Brazil, 2012.

[2] J.G. Speight, Handbook of Petroleum Product Analysis, Wiley-Interscience, Hoboken, NJ, 2002.

[3] G. Totten, S. Westbrook, R. Shah, Fuels and Lubricants Handbook: Technology, Properties, Performance, and Testing, ASTM International, West Conshohocken, PA, 2003. doi:10.1520/mnl37-eb.

[4] J.A. Ardila, F.L.F. Soares, M.A. dos Santos Farias, R.L. Carneiro, Characterization of Gasoline by Raman Spectroscopy with Chemometric Analysis, Anal. Lett. 50 (2017) 1126–1138. doi:10.1080/00032719.2016.1210616.

[5] H.L. Fernandes, I.M.R. Jr, C. Pasquini, J.J.R. Rohwedder, Simultaneous determination of methanol and ethanol in gasoline using NIR spectroscopy: Effect of gasoline composition, Talanta. 75 (2008) 804–810. doi:10.1016/j.talanta.2007.12.025.

[6] U.S. Energy Information Administration, Finished Motor Gasoline Supplied, (2019). https://www.eia.gov/dnav/pet/pet_cons_psup_a_EPM0F_VPP_mbbl_m.htm (accessed August 2, 2019).

[7] ASTM D2699-19, Standard Test Method for Research Octane Number of Spark-Ignition Engine Fuel, in: ASTM Int., 2019. doi:10.1520/D2699-19.

[8] ASTM D2700-19, Standard Test Method for Motor Octane Number of Spark-Ignition Engine Fuel, in: ASTM Int., 2019. doi:10.1520/D2700-19.

[9] ASTM D1319-18, Standard Test Method for Hydrocarbon Types in Liquid Petroleum Products by Fluorescent Indicator Adsorption, in: ASTM Int., 2018. doi:10.1520/D1319-18.

[10] ASTM D4052-18a, Standard Test Method for Density, Relative Density, and API Gravity of Liquids by Digital Density Meter, in: ASTM Int., 2018. doi:10.1520/D4052-18A.

[11] G. Mendes, H.G. Aleme, P.J.S. Barbeira, Determination of octane numbers in gasoline by distillation curves and partial least squares regression, Fuel. 97 (2012) 131–136. doi:10.1016/j.fuel.2012.01.058.

[12] J. Beens, H.T. Feuerhelm, J.C. Fröhling, J. Watt, G. Schaatsbergen, A Comparison of Ten Different Methods for the Analysis of Saturates, Olefins, Benzene, Total Aromatics, and Oxygenates in Finished Gasolines, J. Chromatogr. Sci. 41 (2003) 564–569. doi:10.1093/chromsci/41.10.564.

[13] F.S.L. Costa, R.H.P. Pedroza, D.L. Porto, M.V.P. Amorimb, K.M.G. Lima, Multivariate control charts for simultaneous quality monitoring of isoniazid and rifampicin in a pharmaceutical formulation using a portable near infrared spectrometer, J. Braz. Chem. Soc. 26 (2015) 64–73. doi:10.5935/0103-5053.20140214.

[14] R. da Silva Fernandes, F.S.L. da Costa, P. Valderrama, P.H. Março, K.M.G. de Lima, Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods, J. Pharm. Biomed. Anal. (2012). doi:10.1016/j.jpba.2012.03.004.

[15] P.H. Ciza, P.Y. Sacre, C. Waffo, L. Coïc, H. Avohou, J.K. Mbinze, R. Ngono, R.D. Marini, P. Hubert, E. Ziemons, Comparing the qualitative performances of handheld NIR and Raman spectrophotometers for the detection of falsified pharmaceutical products, Talanta. (2019). doi:10.1016/j.talanta.2019.04.049.

[16] B.W. Wabuyele, S. Sotthivirat, G.X. Zhou, J. Ash, S.S. Dhareshwar, Dispersive Raman Spectroscopy for Quantifying Amorphous Drug Content in Intact Tablets, J. Pharm. Sci. (2017). doi:10.1016/j.xphs.2016.10.014.

[17] R.C. Costa, V.H. Uchida, T.B. Veríssimo Miguel, M.M.L. Duarte, K.M.G. Lima, Quantification of quality parameters in castanhola fruits by NIRS for the development of prediction models using PLS and variable selection algorithms on a laboratory scale, Anal. Methods. (2017). doi:10.1039/c6ay02454h.

[18] E. Guzmán, V. Baeten, J.A.F. Pierna, J.A. García-Mesa, A portable Raman sensor for the rapid discrimination of olives according to fruit quality, Talanta. (2012). doi:10.1016/j.talanta.2012.01.053.

[19] P.I.C. Richardson, H. Muhamadali, D.I. Ellis, R. Goodacre, Rapid quantification of the adulteration of fresh coconut water by dilution and sugars using Raman spectroscopy and chemometrics, Food Chem. (2019). doi:10.1016/j.foodchem.2018.08.038.

[20] K. De Sá Oliveira, L. De Souza Callegaro, R. Stephani, M.R. Almeida, L.F.C. De Oliveira, Analysis of spreadable cheese by Raman spectroscopy and chemometric tools, Food Chem. (2016). doi:10.1016/j.foodchem.2015.08.039.

[21] L. Mandrile, G. Zeppa, A.M. Giovannozzi, A.M. Rossi, Controlling protected designation of origin of wine by Raman spectroscopy, Food Chem. (2016). doi:10.1016/j.foodchem.2016.05.011.

[22] R. Ríos-Reina, D.L. García-González, R.M. Callejón, J.M. Amigo, NIR spectroscopy and chemometrics for the typification of Spanish wine vinegars with a protected designation of origin, Food Control. (2018). doi:10.1016/j.foodcont.2018.01.031.

[23] A. Pérez, Y.A. Prada, R. Cabanzo, C.I. González, E. Mejía-Ospino, Diagnosis of chagas disease from human blood serum using surface-enhanced Raman scattering (SERS) spectroscopy and chemometric methods, Sens. Bio-Sensing Res. (2018). doi:10.1016/j.sbsr.2018.10.003.

[24] J. V. Rau, F. Marini, M. Fosca, C. Cippitelli, M. Rocchia, A. Di Napoli, Raman spectroscopy discriminates malignant follicular lymphoma from benign follicular hyperplasia and from tumour metastasis, Talanta. (2019). doi:10.1016/j.talanta.2018.10.086.

[25] J.C.L. Alves, R.J. Poppi, Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration, Fuel. (2016). doi:10.1016/j.fuel.2015.10.079.

[26] R.H. de Paula Pedroza, J.T.N. Nicácio, B.S. dos Santos, K.M.G. de Lima, Determining the Kinematic Viscosity of Lubricant Oils for Gear Motors by Using the Near Infrared Spectroscopy (NIRS) and the Wavelength Selection, Anal. Lett. 46 (2013) 1145–1154. doi:10.1080/00032719.2012.751542.

[27] V.O. Santos, F.C.C. Oliveira, D.G. Lima, A.C. Petry, E. Garcia, P.A.Z. Suarez, J.C. Rubim, A comparative study of diesel analysis by FTIR, FTNIR and FT-Raman spectroscopy using PLS and artificial neural network analysis, Anal. Chim. Acta. (2005). doi:10.1016/j.aca.2005.05.042.

[28] C.K. Muro, K.C. Doty, L. de Souza Fernandes, I.K. Lednev, Forensic body fluid identification and differentiation by Raman spectroscopy, Forensic Chem. (2016). doi:10.1016/j.forc.2016.06.003.

[29] M.R. de Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, R.J. Poppi, Discrimination between authentic and counterfeit banknotes using raman spectroscopy and PLS-DA with uncertainty estimation, Microchem. J. 109 (2013) 170–177. doi:10.1016/j.microc.2012.03.006.

[30] T.M. Barbosa, L.A.S. de Lima, M.C.D. dos Santos, S.D. Vasconcelos, R.A. Gama, K.M.G. Lima, A novel use of infra-red spectroscopy (NIRS and ATR-FTIR) coupled with variable selection algorithms for the identification of insect species (Diptera: Sarcophagidae) of medico-legal relevance, Acta Trop. (2018). doi:10.1016/j.actatropica.2018.04.025.

[31] O.M.D. Lutz, G.K. Bonn, B.M. Rode, C.W. Huck, Reproducible quantification of ethanol in gasoline via a customized mobile near-infrared spectrometer, Anal. Chim. Acta. 826 (2014) 61–68. doi:10.1016/j.aca.2014.04.002.

[32] S.J. Choquette, S.N. Chesler, D.L. Duewer, S. Wang, T.C. O'Haver, Identification and Quantitation of Oxygenates in Gasoline Ampules Using Fourier Transform Near-Infrared and Fourier Transform Raman Spectroscopy, Anal. Chem. 68 (1996) 3525–3533. doi:10.1021/ac960451v.

[33] R.M. Correia, E. Domingos, V.M. Cáo, B.R.F. Araujo, S. Sena, L.U. Pinheiro, A.M. Fontes, L.F.M. Aquino, E.C. Ferreira, P.R. Filgueiras, W. Romão, Portable near infrared spectroscopy applied to fuel quality control, Talanta. 176 (2018) 26–33. doi:10.1016/j.talanta.2017.07.094.

[34] F.M. Fortunato, A.L. Vieira, J.A. Gomes Neto, G.L. Donati, B.T. Jones, Expanding the potentialities of standard dilution analysis: Determination of ethanol in gasoline by Raman spectroscopy, Microchem. J. 133 (2017) 76–80. doi:10.1016/j.microc.2017.03.015.

[35] Q. Ye, Q. Xu, Y. Yu, R. Qu, Z. Fang, Rapid and quantitative detection of ethanol proportion in ethanol-gasoline mixtures by Raman spectroscopy, Opt. Commun. 282 (2009) 3785–3788. doi:10.1016/j.optcom.2009.06.034.

[36] A.C. de M. Bezerra, D. de O. Silva, G.H.M. de Matos, J.P. dos Santos, C.N. Borges, L. Silveira, M.T.T. Pacheco, Quantification of anhydrous ethanol and detection of adulterants in commercial Brazilian gasoline by Raman spectroscopy, Instrum. Sci. Technol. 47 (2019) 90–106. doi:10.1080/10739149.2018.1470535.

[37] L.S. Mendes, F.C.C. Oliveira, P.A.Z. Suarez, J.C. Rubim, Determination of ethanol in fuel ethanol and beverages by Fourier transform (FT)-near infrared and FT-Raman spectrometries, Anal. Chim. Acta. 493 (2003) 219–231. doi:10.1016/S0003-2670(03)00870-5.

[38] M. V. Reboucas, J.B. Santos, M.F. Pimentel, L.S.G. Teixeira, A novel approach for development of a multivariate calibration model using a Doehlert experimental design: Application for prediction of key gasoline properties by Near-infrared Spectroscopy, Chemom. Intell. Lab. Syst. 107 (2011) 185–193. doi:10.1016/j.chemolab.2011.03.007.

[39] R.M. Balabin, R.Z. Safieva, E.I. Lomakina, Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction, Chemom. Intell. Lab. Syst. 88 (2007) 183–188. doi:10.1016/j.chemolab.2007.04.006.

[40] N.C. da Silva, C.J. Cavalcanti, F.A. Honorato, J.M. Amigo, M.F. Pimentel, Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters, Anal. Chim. Acta. 954 (2017) 32–42. doi:10.1016/j.aca.2016.12.018.

[41] J.B. Cooper, K.L. Wise, W.T. Welch, M.B. Sumner, B.K. Wilt, R.R. Bledsoe, Comparison of Near-IR, Raman, and Mid-IR Spectroscopies for the Determination of BTEX in Petroleum Fuels, 51 (1997) 1613–1620.

[42] A.F. Parisi, L. Nogueiras, H. Prieto, On-line determination of fuel quality parameters using near-infrared spectrometry with fibre optics and multivariate calibration, Anal. Chim. Acta. 238 (1990) 95–100. doi:10.1016/S0003-2670(00)80527-9.

[43] P.E. Flecher, W.T. Welch, S. Albin, J.B. Cooper, Determination of octane numbers and Reid vapor pressure in commercial gasoline using dispersive fiber-optic Raman spectroscopy, Spectrochim. Acta - Part A Mol. Spectrosc. 53 (1997) 199–206.

[44] J.B. Cooper, K.L. Wise, J. Groves, W.T. Welch, Determination of Octane Numbers and Reid Vapor Pressure of Commercial Petroleum Fuels Using FT-Raman Spectroscopy and Partial Least-Squares Regression Analysis, Anal. Chem. 67 (1995) 4096–4100. doi:10.1021/ac00118a011.

[45] M. Voigt, R. Legner, S. Haefner, A. Friesen, A. Wirtz, M. Jaeger, Using fieldable spectrometers and chemometric methods to determine RON of gasoline from petrol stations: A comparison of low-field 1H NMR@80 MHz, handheld RAMAN and benchtop NIR, Fuel. 236 (2019) 829–835. doi:10.1016/j.fuel.2018.09.006.

[46] R.R. de Oliveira, R.H.P. Pedroza, A.O. Sousa, K.M.G. Lima, A. de Juan, Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy, Anal. Chim. Acta. 985 (2017) 41–53. doi:10.1016/j.aca.2017.07.038.

[47] X. Zhang, X. Qi, M. Zou, J. Wu, Rapid detection of gasoline by a portable Raman spectrometer and chemometrics, J. Raman Spectrosc. 43 (2012) 1487–1491. doi:10.1002/jrs.4076.

[48] S. Ahuja, N. Jespersen, Modern Instrumental Analysis, Elsevier Science, New York, NY, 2006.

[49] Douglas A. Skoog, F.J. Holler, S.R. Crouch, Principles of Instrumental Analysis, Cengage Learning, Boston, MA, 2016.

[50] J. Workman, Jr., L. Weyer, Practical Guide to Interpretive Near-Infrared Spectroscopy, CRC Press, Boca Raton, FL, 2007. doi:10.1201/9781420018318.

[51] S. Hansen, S. Pedersen-Bjergaard, K. Rasmussen, Introduction to Pharmaceutical Analysis, John Wiley & Sons, Hoboken, NJ, 2011. doi:10.1002/9781119953647.ch1.

[52] P. Larkin, Infrared and Raman Spectroscopy; Principles and Spectral Interpretation, Elsevier, Waltham, MA, 2011. doi:10.1016/C2010-0-68479-3.

[53] P. Vandenabeele, Practical Raman Spectroscopy - An Introduction, John Wiley & Sons, Hoboken, NJ, 2013. doi:10.1002/9781119961284.

[54] R.G. Brereton, Chemometrics for Pattern Recognition, John Wiley & Sons, Hoboken, NJ, 2009. doi:10.1002/9780470746462.

[55] S. Brown, R. Tauler, B. Walczak, Comprehensive Chemometrics, Elsevier B.V., Amsterdam, Netherlands, 2009. doi:10.1016/C2009-0-28356-5.

[56] T. Naes, A User Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK, 2002.

[57] K. Varmuza, P. Filzmoser, Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, FL, 2009.

[58] A.C. Olivieri, Introduction to Multivariate Calibration - A Practical Approach, Springer, New York, NY, 2018. doi:10.1007/978-3-319-97097-4.

[59] S. Chountasis, V.N. Katsikis, D. Pappas, A. Perperoglou, The whittaker smoother and the moore-penrose inverse in signal reconstruction, Appl. Math. Sci. 6 (2012) 1205–1219.

[60] P.H.C. Eilers, H.F.M. Boelens, Baseline Correction with Asymmetric Least Squares Smoothing, Life Sci. (2005) 1–26. doi:10.1021/ac034173t.

[61] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. doi:10.1016/0169-7439(87)80084-9.

[62] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. (1986). doi:10.1016/0003-2670(86)80028-9.

[63] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/C3AY41907J.

[64] D.W. Sun, Infrared Spectroscopy for Food Quality Analysis and Control, Elsevier Inc., Burlington, MA, 2009. doi:10.1016/B978-0-12-374136-3.X0001-6.

[65] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods. 5 (2013) 3790. doi:10.1039/c3ay40582f.

[66] J. Kuligowski, D. Pérez-guaita, G. Quintás, Statistical Analysis in Proteomics, 1362 (2016) 175–184. doi:10.1007/978-1-4939-3106-4.

[67] L. V. Madden, P.A. Paul, Assessing heterogeneity in the relationship between wheat yield and Fusarium head blight intensity using random-coefficient mixed models, Phytopathology. 99 (2009) 850–860. doi:10.1094/PHYTO-99-7-0850.

[68] O. Schabenberger, F.J. Pierce, Contemporary statistical models for the plant and soil sciences, 2001.

[69] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemom. Intell. Lab. Syst. 174 (2018) 33–44. doi:10.1016/j.chemolab.2017.12.004.

[70] A.G. González, M.A. Herrador, A.G. Asuero, Intra-laboratory testing of method accuracy from recovery assays, Talanta. 48 (1999) 729–736. doi:10.1016/S0039-9140(98)00271-9.

[71] M. Vosough, S.N. Eshlaghi, R. Zadmard, On the performance of multiway methods for simultaneous quantification of two fluoroquinolones in urine samples by fluorescence spectroscopy and second-order calibration strategies, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 136 (2015) 618–624. doi:10.1016/j.saa.2014.09.075.

[72] CRC Fuels for Advanced Combustion Engines Working Group (FACE) | Coordinating Research Council, (n.d.). https://crcao.org/face/ (accessed September 10, 2019).

[73] J. Burri, R. Crockett, R. Hany, D. Rentsch, Gasoline composition determined by 1 H NMR spectroscopy, 83 (2004) 187–193. doi:10.1016/S0016-2361(03)00261-8.

[74] C.R. Kaiser, J.L. Borges, R. Anderson, D.A. Azevedo, L.A.D. Avila, Quality control of gasoline by 1 H NMR : Aromatics , olefinics , paraffinics , and oxygenated and benzene contents, Fuel. 89 (2010) 99–104. doi:10.1016/j.fuel.2009.06.023.

[75] H.W. Siesler, Y. Ozaki, S. Kawata, Near-Infrared Spectroscopy. Principles, Instruments, Applications, 2002. doi:10.1002/cem.762.

[76] K.M. Tan, I. Barman, N.C. Dingari, G.P. Singh, T.F. Chia, W.L. Tok, Toward the development of Raman spectroscopy as a nonperturbative online monitoring tool for gasoline adulteration, Anal. Chem. 85 (2013) 1846–1851. doi:10.1021/ac3032349.

[77] S. Corsetti, D. McGloin, J. Kiefer, Comparison of Raman and IR spectroscopy for quantitative analysis of gasoline/ethanol blends, Fuel. 166 (2016) 488–494. doi:10.1016/j.fuel.2015.11.018.

[78] F.S. De Oliveira, L.S.G. Teixeira, M.C.U. Araujo, M. Korn, Screening analysis to detect adulterations in Brazilian gasoline samples using distillation curves, Fuel. 83 (2004) 917–923. doi:10.1016/j.fuel.2003.09.018.

[79] D.L. Flumignan, F. de Oliveira Ferreira, A.G. Tininis, J.E. de Oliveira, Multivariate calibrations in gas chromatographic profiles for prediction of several physicochemical parameters of Brazilian commercial gasoline, Chemom. Intell. Lab. Syst. 92 (2008) 53–60. doi:10.1016/j.chemolab.2007.12.003.

[80] L. Antonio, F. De Godoy, M.P. Pedroso, E.C. Ferreira, F. Augusto, R.J. Poppi, Prediction of the physicochemical properties of gasoline by comprehensive two-dimensional gas chromatography and multivariate data processing, J. Chromatogr. A. 1218 (2011) 1663–1667. doi:10.1016/j.chroma.2011.01.056.

[81] S.J. Swarin, C.A. Drumm, Prediction of Gasoline Properties with NearInfrared Spectroscopy and Chemometrics, SAE Tech. Pap. Ser. 1 (1991). doi:10.4271/912390.

[82] J.M. De Paulo, J.E.M. Barros, P.J.S. Barbeira, A PLS regression model using flame spectroscopy emission for determination of octane numbers in gasoline, Fuel. 176 (2016) 216–221. doi:10.1016/j.fuel.2016.02.033.

[83] M. Wahadoszamen, A. Rahaman, N.M.R. Hoque, A. I Talukder, K.M. Abedin, A.F.M.Y. Haider, Laser Raman spectroscopy with different excitation sources and extension to surface enhanced raman spectroscopy, J. Spectrosc. 2015 (2015). doi:10.1155/2015/895317.