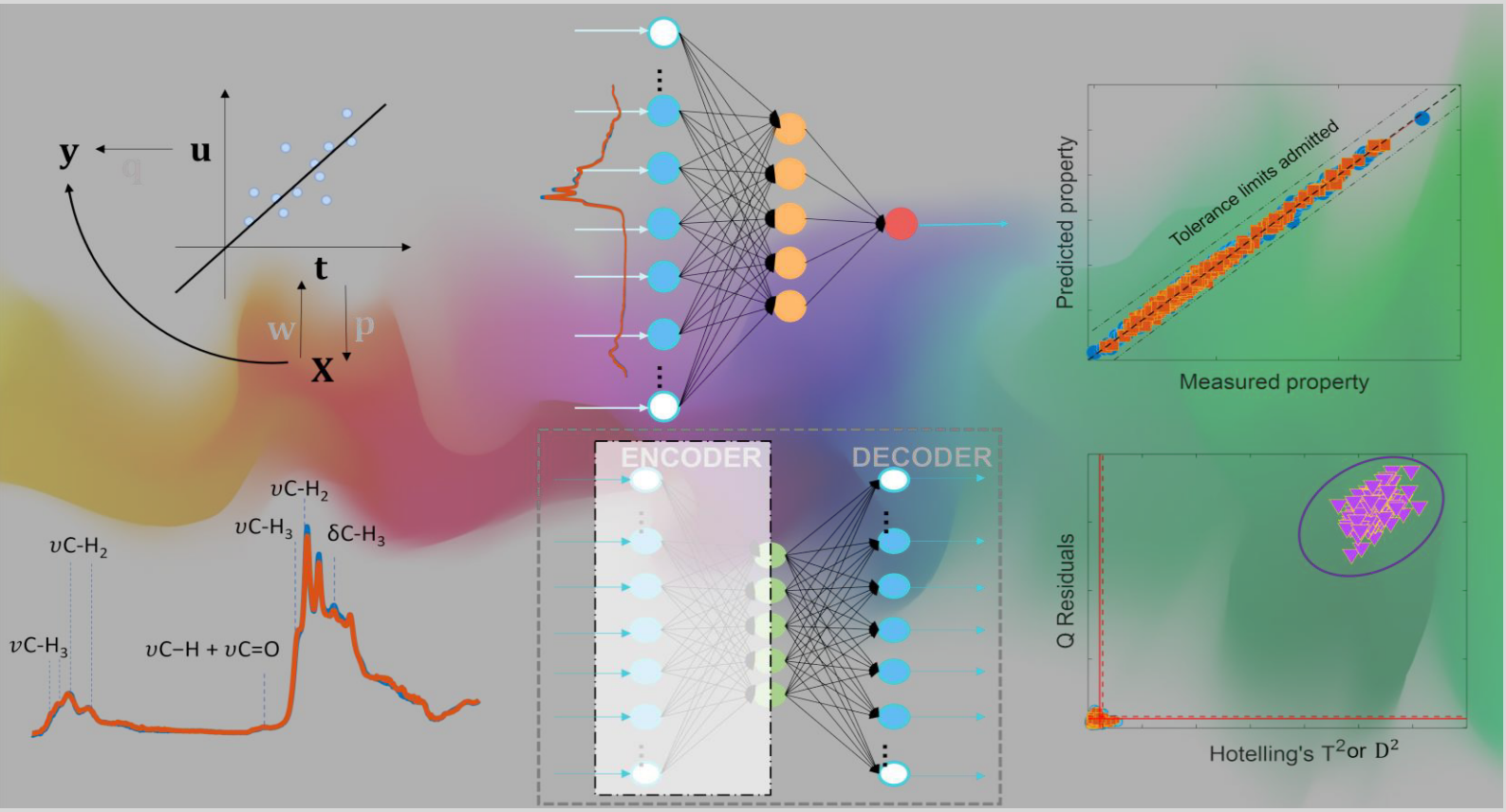
DETERMINATION OF DIESEL PROPERTIES BY INFRARED SPECTROSCOPY AND MULTIVARIATE CALIBRATION

María Suliany Rodríguez Barrios

UNIVERSITAT ROVIRA i VIRGILI

Determination of diesel properties by infrared spectroscopy and multivariate calibration

MARÍA SULIANY RODRÍGUEZ BARRIOS



DOCTORAL THESIS
2024

María Suliany Rodríguez Barrios

Determination of diesel properties by infrared spectroscopy and multivariate calibration

Doctoral Thesis

Supervised by

Dr. Joan Ferré

Dr. Maria Soledad Larrechi

Departament de Química Analítica i Química Orgànica

UNIVERSITAT ROVIRA i VIRGILI

Tarragona

2024

UNIVERSITAT ROVIRA i VIRGILI
Dept. de Química Analítica
i Química Orgànica

WE STATE that the present study, entitled “Determination of diesel properties by infrared spectroscopy and multivariate calibration”, presented by María Suliany Rodríguez Barrios for the award of the degree of Doctor, has been carried out under our supervision at the Department of Analytical Chemistry and Organic Chemistry of the Universitat Rovira i Virgili (URV).

Tarragona, May 6th, 2024

Doctoral Thesis Supervisors

Dr. Joan Ferré Baldrich

Dr. Maria Soledad Larrechi García

UNIVERSITAT ROVIRA i VIRGILI

The work performed in the present Doctoral Thesis has been possible thanks to the Martí i Franquès Research Fellowship Program (2021PMF-PIPF-1) and the funding of Repsol Petróleo S.A.(T19221S) and the Agency for Management of University and Research Grants (AGAUR, 2021-SGR00705).

Acknowledgments

First of all, I would like to express my deep gratitude and genuine acknowledgment to my supervisors, Prof. Maria Soledad Larrechi García and Prof. Joan Ferré, for granting me the privilege to pursue this doctoral thesis. This unique opportunity has been crucial for consolidating my academic training and also for my personal growth. Thank you for your invaluable assistance throughout the supervision of my research, for your dedication, patience, and consideration, and for caring not only about my well-being at the university but also about aspects of the daily life of an international student. Thank you for everything I have learned from you.

I would also like to thank Enric Ruiz Morillas from Repsol Petróleo, Tarragona, for his invaluable assistance in the experimental part of this thesis, his availability, and his thoughtful comments on the work carried out.

I would like to thank Dr. Santiago Macho for his help in the early stages of this thesis. I would like to thank the research group members who have been aware of me, especially Prof. Itziar Ruisánchez.

I am deeply grateful to Professor Alberto Coronas, who gave me the opportunity to complete my master's studies, for his valuable advice and constant motivation.

Thanks to my friends for always being aware of the development of this thesis and to my colleagues from URV for the shared moments.

I would like to express my sincere gratitude to the entire Xino-Xano family, especially Ángeles, for their lovely care of my daughter. I have felt calm on overwhelming days, knowing that she was in such loving hands.

A special thank you to my parents and my sister for their unconditional love despite the distance. This thesis is the culmination of many years of study in which your constant support has been essential. Every milestone in my academic journey owes much to your encouragement, and I am delighted to celebrate these achievements with you. I would also like to thank my entire family for their motivation, particularly my grandparents and my mother-in-law.

I would like to thank my daughter and my husband, my main source of inspiration and motivation, for their infinite love. I am deeply grateful to my husband, who has supported me during all these challenging years, for encouraging me to persevere and finish this stage hundreds of times, believing in me more than myself, and for his wise advice. Without his tireless efforts, this work would not have been possible.

Abbreviations

AD: applicability domain

AE: autoencoder

ANN: artificial neural network

ASTM: ASTM International

ATR-FTIR: attenuated total reflectance-Fourier transform infrared

CFPP: cold filter plugging point

CN: cetane number

CORES: Strategic Petroleum Products Reserves Corporation

CSMWPLS: changeable size moving window partial least squares

CT: calibration transfer

di-PLS: domain invariant PLS regression

DOP: dynamic orthogonal projection

DS: direct standardization

D^2: squared Mahalanobis distance

EN: european norm

FAME: fatty acid methyl esters

FFNN: feed-forward neural network

FP: flash point

FTIR: Fourier transform infrared

GLSW: generalized least squares weighting

HATR: horizontal attenuated total reflectance (HATR)

HDS: hydrodesulfuration

IR: Infrared

ISO: International Organization for Standardization

K-ANN: Kohonen artificial neural network

KS: Kennard Stone

LOOCV: leave-one-out cross-validation

LVs: latent variables

MC: mean centering

MCSMWPLSR: modified changeable size moving window partial least squares

MIR: mid-infrared

MLR: multiple linear regression

MSC: multiplicative scatter correction

MSPC: multivariate statistical process control

MU: model updating

MWPLSR: moving window partial least squares regression

NIPALS: non-iterative partial least squares

NIR: near-infrared

NMR: nuclear magnetic resonance

OECD: Organisation for Economic Cooperation and Development

OSC: orthogonal signal correction

PCA: principal component analysis

PCR: principal component regression,

PDS: piecewise direct standardization

PLSR: partial least squares regression

QSAR: Quantitative-structure activity relationship

r: correlation coefficient

Ref: reference

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

RS: reverse standardization

R_c^2: coefficient of determination of calibration

R_{cv}^2: coefficient of determination of cross-validation

R_p^2: coefficient of determination of prediction

SBCP: slope and bias correction

SCA: scatter component analysis

SCMWPLS: searching combination moving window partial least squares

SNV: standard normal variate

SPE: standard prediction error

SVM: support vector machines

SVR: support vector regression

TCA: transfer component analysis

t-SNE: T-distributed stochastic neighbor embedding

UNE: Spanish Standardization Association

Q: spectral residuals

Summary

The potential of infrared (IR) spectroscopy combined with multivariate calibration has been well-vetted over the years for its ability to predict the quality parameters of fuel samples. The main objective of this doctoral thesis is to develop calibration models based on IR spectroscopy to predict diesel quality properties. These models are of great interest in the petrochemical industry since they can be used in offline (laboratory) and online (process) analysis as an alternative to standard methods.

In this work, a broad set of diesel samples collected at the Repsol refinery in Tarragona (Spain) during almost four years (2018-2022) from two streams (namely, desulfurized diesel (stream 1) and commercial diesel (stream 2)) were considered to predict eleven properties of interest. These properties are: (1) distillation temperatures at 65%, (2) 85%, and (3) 95% recovered (T85%, T85%, and T95%), (4) flash point, (5) cloud point, (6) density, (7) cetane number, (8) sulfur content, (9) viscosity, (10) cold filter plugging point (CFPP), and (11) fatty acid methyl esters (FAME).

The physicochemical properties of a diesel fuel emerge from its chemical composition, which also defines the characteristics of its IR spectrum. We have characterized the desulfurized and commercial samples in terms of their physicochemical properties and spectral features using analysis of correlations between properties and multivariate analysis of the measured spectra, respectively. From the results, we concluded that: 1) except for the distillation temperatures in commercial samples, the properties were not significantly correlated; 2) Three spectral clusters were identified: desulfurized samples, commercial samples containing FAME, and commercial samples without FAME content; 3) The difference in the property values of the samples seems to be more influenced by the origin of the samples (from streams 1 or 2) rather than by the FAME content. This latter observation has led to the development of specific calibration models for samples from each stream.

Partial least squares regression (PLS) and artificial neural networks (ANN) calibration models have been considered to model IR spectrum-property relationships. For the ANN approach, a new strategy for establishing the limits of the applicability domain has been developed. Two limits based on 1) the 0.99 quantile of the squared Mahalanobis distance calculated from the network activations of the training set and 2) the 0.99 quantile of the reconstruction error of the training spectra using either an autoencoder network or a decoder network were considered.

Using PLS and ANN calibration models, we have accurately predicted the density and cetane number of both desulfurized and commercial samples and the FAME content and viscosity of commercial samples. Only the models for determining density and FAME content could be considered valid for diesel routine analysis. Comparatively, the predictive performances of the ANN models were superior for density, T65%, T85%, flash point, CFPP, and FAME content in commercial samples, and T95%, flash point, cloud point, sulfur content in desulfurized samples. In turn, the PLS model outperformed the ANN model in predicting the density of samples from stream 1 and T95% and cloud point of commercial samples. The predictive ability of ANN and PLS models was similar for the remaining properties.

Monitoring the predictive ability of the PLS models over ten months showed that the prediction errors did not vary significantly for most of the properties with the exception of flash point (commercial samples), sulfur content (desulfurized samples), and FAME content. However, despite the observed decline in the predictive performance of the PLS model for FAME content during this monitoring period, it remained the only valid model for use in routine diesel analysis.

The deterioration of the predictive ability of the PLS model to predict density was confirmed when it was applied to spectra measured on an infrared spectrophotometer different from the one used to develop the model. Thus, three model adaptations from the developed PLS model were compared: 1) domain invariant partial least squares (di-PLS), 2) dynamic orthogonal projection (DOP), and 3) model updating (MU). The results showed 80%, 87%, and 92% reductions in root mean squared error prediction (RMSEP) after applying di-PLS, DOP, and MU, respectively. The results point to MU as the only valid model adaptation to be used for control in diesel analysis.

Index

List of Figures

List of Tables

Chapter I
Introduction, objectives, and structure

Chapter I

I.1 Introduction

I.1.1 Diesel relevance. Production process

Diesel is one of the most widely used fuels around the world. It has a crucial role in the global economy and in raising living standards as it is used in internal combustion engines of automotive vehicles in every mode of transport (maritime, rail, air, and road), electricity generation, and domestic heating, among others. In European OECD countries, the production of diesel from crude oil represented the most significant share (33.3%) of the average production of refineries in 2022 (Figure I-1) [1]. Specifically in Spain, diesel consumption in 2022 represented 55% of the total consumption of petroleum products, as reported by the Strategic Petroleum Products Reserves Corporation (CORES, by its acronym in spanish) [2].

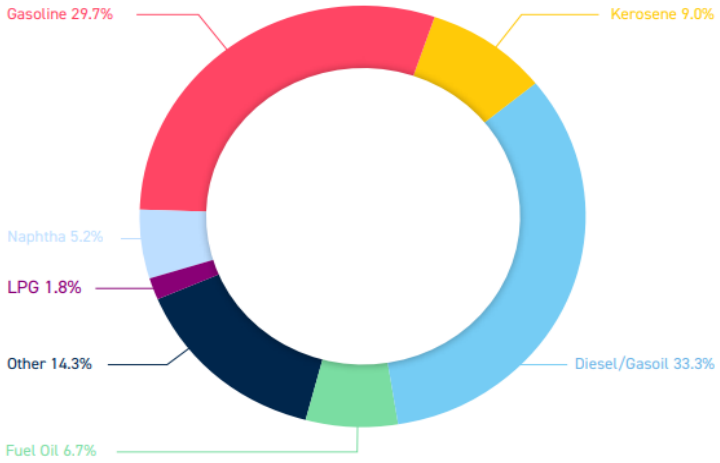


Figure I-1. Average refinery output by product type in OECD Europe in 2022. Adapted from the International Energy Agency [1].

In the last 20 years, the change in consumption from gasoline to diesel has been due to, among other factors, efficiency, energy density, and lower CO_2 emissions of diesel as compared to gasoline. This has led to a strong global increase in demand for diesel. In this regard, it is estimated that by 2045, the demand for diesel will amount to approximately 29.6 million barrels per day, which represents an increase of 3.30 million barrels per day more than in 2020 [3].

Diesel or petrodiesel is a mixture of various fractions obtained during petroleum refining [4]. Specifically, it comes from mixtures of the medium and heavy fractions of the direct distillation of crude oil [5], which are treated to improve their quality and then mixed to obtain the final product with the required specifications. In some cases, other products (biodiesel, for instance) that do not come from petroleum are added to the final mixture [6].

Figure I-2 illustrates diesel production in a refinery, where diesel is obtained from mixtures of five streams. Stream **A** comes from the middle fractions of the direct distillation of crude oil (atmospheric distillation). Stream **B** is obtained from the hydrotreatment of the mixture of the heavy fractions of this direct distillation and other streams. Hydrotreatment is a hydrogenation process where hydrogen replaces the heteroatoms (nitrogen, oxygen, and sulfur) present in the hydrocarbons and binds with them to form ammonia (NH_3), water (H_2O), and hydrogen sulfide (H_2S) [5,6]. Also, the saturation of aromatic compounds and olefins is produced [7]. Stream **C** comes from the hydrodesulfurization process (ISOMAX unit) of the lighter fractions obtained from vacuum distillation (*Light Vacuum Gasoil*). This process greatly improves the quality of the diesel by decreasing the concentration of sulfur, thus reducing corrosivity, environmentally damaging emissions, and engine wear caused by sulfur compounds. Stream **D** is obtained during hydrocracking of the heavy fractions of vacuum distillation (GOPV). Hydrocracking is a catalytic cracking process that allows the conversion of heavy hydrocarbons into lighter ones by incorporating hydrogen under pressure in the presence of a catalyst [5,8]. This process is also used to reduce the sulfur and nitrogen content. Finally, stream **E** comes from the visbreaking process of the residue from vacuum distillation, which is subsequently hydrodesulfurized. Visbreaking reduces the viscosity of heavy residues from distillations.

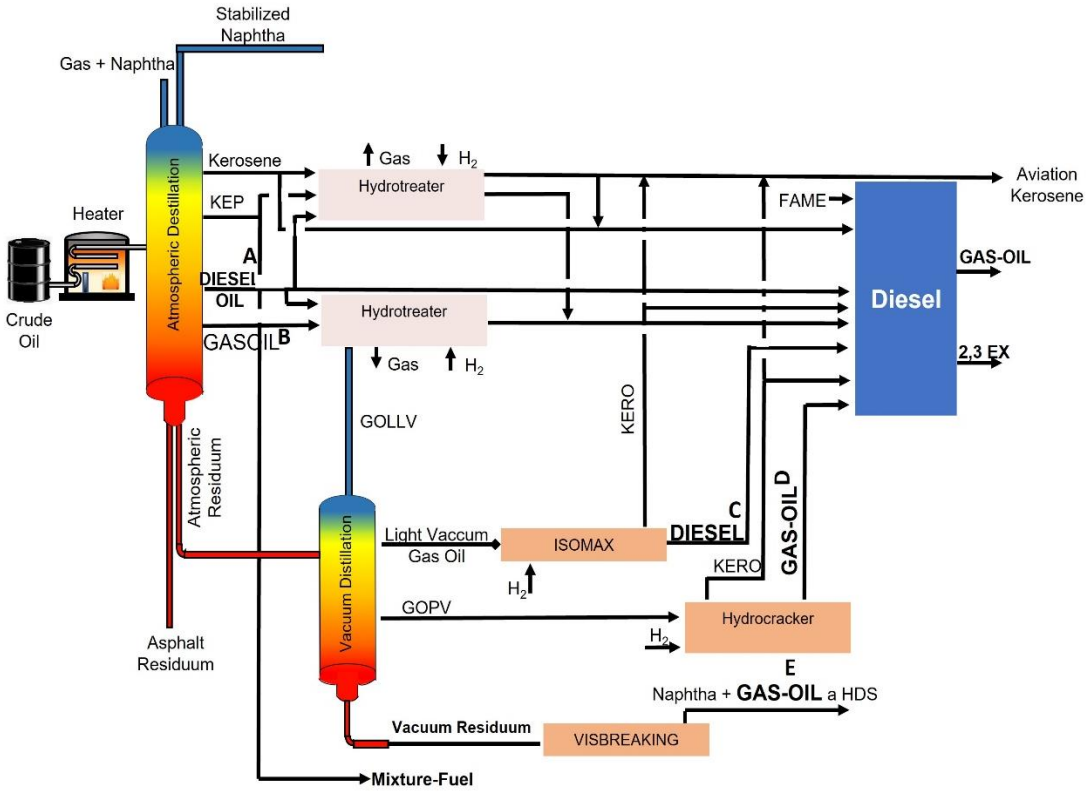


Figure I-2. Simplified outline of the diesel production process.

I.1.2 Chemical composition and physicochemical properties of diesel

Diesel is made up of paraffinic, naphthenic, and aromatic compounds. Their carbon chains can include between 8 and 22 carbon atoms, although most have between 17 and 20 carbon atoms and are paraffinic chains. Table I-1 shows the most representative hydrocarbons of diesel. The different proportions of paraffin, naphthenes, and aromatics determine the quality of the final product. It is worth mentioning that other chemical compounds, such as sulfur, nitrogen, and oxygen, may be present in small amounts in diesel and play an essential role in its properties [9].

The different relative proportions of diesel hydrocarbons determine the differences between diesel types. For example, one can expect that the higher the content of paraffinic compounds, the higher the value of the cetane number, and conversely, the higher the aromatics content, the lower the cetane value [10]. A low flash point may also indicate contamination of the diesel sample with gasoline. At the same time, high-density values are related to an increase in the content of fatty acid methyl esters (FAME), an increase in un-saturations, or an increase in hydrocarbons with a shorter chain length [11–13].

Determining the physicochemical properties of diesel is crucial for safety and regulatory purposes. Traditional characterization procedures are standardized by official organizations such as the American Society for Testing and Materials (ASTM), European Norms (EN), the Spanish Standardization Association (UNE), and the International Organization for Standardization (ISO). These procedures involve various analytical determinations that are usually carried out off-line in chemical laboratories. Among them, the determination of density, sulfur, and FAME content are of prime importance since they are parameters that have a substantial impact on the price and quality of diesel. Table I-2 summarizes the main properties that are used to characterize the quality of diesel, the analytical methods recognized by the official organizations, the techniques involved in their determination, and a brief description of the property. The precision data of the reference analytical method for each property, including the repeatability and reproducibility and standard deviation of reproducibility (sR), are listed in Table I-3.

Determining diesel specifications using standard methods has limitations since they require specific instrumentation that must be purchased and maintained, are time-consuming, or require intensive sample handling so that the cost-per-analysis can be high. One case is the standard method of determining the cetane number, which requires a large sample volume and high-purity hydrocarbon references. In addition, due to the time needed for a test, the method is not feasible for on-line control of diesel blending. Quick

and cost-effective assessments are preferred in the context of quality control and process control.

Table I-1. The most representative hydrocarbons of diesel, their molecular formula, and the corresponding hydrocarbon group [9].

Hydrocarbon class	Compound	Molecular formula
n-Paraffins	Decane	$C_{10}H_{22}$
	n-Pentadecane	$C_{15}H_{32}$
	n-Hexadecane (Cetane)	$C_{16}H_{34}$
	Eicosane	$C_{20}H_{42}$
Isoparaffins	2,2-Dimethyloctane	$C_{10}H_{22}$
	3-Ethyldecane	$C_{12}H_{26}$
	4,5-Diethyloctane	$C_{12}H_{26}$
	2-Methyltetradecane	$C_{15}H_{32}$
	Heptamethylnonane	$C_{16}H_{35}$
	8-Propylpentadecane	$C_{18}H_{38}$
	7,8-Diethyltetradecane	$C_{18}H_{38}$
	9,10-Dimethyloctane	$C_{20}H_{40}$
	2-Methylnonadecane	$C_{20}H_{42}$
Naphthenes	cis-Decalin	$C_{10}H_{18}$
	n-Butylcyclohexane	$C_{10}H_{20}$
	n-Pentylcyclopentane	$C_{10}H_{20}$
	3-Cyclohexylhexane	$C_{12}H_{24}$
	n-Nonylcyclohexane	$C_{15}H_{30}$
	n-Decylcyclopentane	$C_{15}H_{30}$
	2-Methyl-3-cyclohexylnonane	$C_{16}H_{32}$
	n-Tetradecylcyclohexane	$C_{20}H_{40}$
	n-Pentadecylcyclopentane	$C_{20}H_{40}$
	2-Cyclohexyltetradecane	$C_{20}H_{40}$
Aromatic	Naphthalene	$C_{10}H_{8}$
	Tetralin	$C_{10}H_{12}$
	1,3-Diethylbenzene	$C_{10}H_{14}$
	1-Methylnaphthalene	$C_{11}H_{10}$
	n-Pentylbenzene	$C_{11}H_{16}$
	Biphenyl	$C_{12}H_{10}$
	Anthracene	$C_{14}H_{10}$
	1-Butylnaphthalene	$C_{14}H_{16}$
	1-Pentylnaphthalene	$C_{15}H_{18}$
	n-Nonylbenzene	$C_{15}H_{24}$
	2-Octylnaphthalene	$C_{18}H_{24}$
	1-Decylnaphthalene	$C_{20}H_{28}$
	n-Tetradecylbenzene	$C_{20}H_{34}$

Table I-2. Physicochemical properties determined in diesel quality routine analysis and the methods used for their determination.

Property	Description	ASTM method	EN 590 method	Method and/or instrumentation
Distillation temperatures at 65%, 85% and 95% recovered	The temperature at which 65%, 85%, and 95% of the diesel has been distilled.	ASTM D2887	EN ISO 3405	Manual or automatic distillation at atmospheric pressure
Flash point	The lowest temperature at which vapors from a diesel sample ignite when exposed to air under controlled laboratory conditions.	ASTM D93	EN ISO 2719	Pensky-Martens Method in closed glass
Cloud point	The temperature below which the wax (paraffin crystals) in diesel begins to solidify, taking on a cloudy appearance.	ASTM D2500 (Manual Method) ASTM D 5772 (Automatic Method)	EN ISO 23015	Constant temperature cooling bath
Density	Mass of diesel per unit volume at 15 °C and 101.3 kPa.	ASTM D4052	-----	Oscillating U-Tube density meter method.
Cetane number	Measurement of the ignition quality of diesel in compression ignition engines under controlled conditions. It is quantified as the volume percentage of cetane (n-hexadecane) in a reference blend which exhibits an ignition delay equivalent to that of the analyzed fuel.	ASTM D613	UNE-EN 16715	CFR Standard test engine method cetane analyzer
Sulfur content	Amount of sulfur present in diesel samples.	ASTM D 5453	EN ISO 20846	Fluorescence in the ultraviolet
Viscosity	Time for a volume of diesel to flow under gravity through viscometer at 40 °C. The kinematic viscosity of diesel indicates its tendency to form deposits in the engine.	ASTM D 445	ISO 3104	Capillary viscometer
Cold filter plugging point (CFPP)	The lowest temperature at which a given volume of diesel fuel still passes through a standardized filtration device at a specific time, when cooled under standardized conditions.	ASTM D6371	UNE-EN-116	Standardized filter device
FAME	Amount of fatty acid methyl esters (FAME), as it is a mixture of esters with different chain lengths and saturation degrees, which can be blended with diesel (up to 20%) for use in diesel engines.	ASTM D7806	UNE-EN 14078	Infrared spectroscopy

Table I-3. Reference methods and their precision.

Property	Method	Repeatability (r)	Reproducibility (R)	sR
T65%	ASTM D 86	1.1	3.6	1.3
T85%	ASTM D 86	1.3	4.8	1.7
T95%	ASTM D 86	2.9	9.0	3.2
Flash Point	ASTM D 93	2	5	2
Cloud Point	ASTM D 5772	1.3	3.3	1.2
Density	ASTM D4052	0.1	0.5	0.2
Cetane number	EN 16715	0.6	1.5	0.5
Sulfur content	EN ISO 20846	1.2	2.4	0.9
Viscosity (μ)	ASTM D 445	0.016	0.031	0.011
Cold filter plugging point	EN-116	1.5	3.6	1.30
FAME	EN 14078	0.2	0.9	0.3

I.1.3 Infrared spectroscopy in diesel analysis

Compared to traditional methods, applying empirical mathematical equations and theoretical models such as the Group Contribution Method can be simple and more cost-effective for predicting some physicochemical properties [11,14–16]. Alternatively, spectroscopic techniques such as vibrational spectroscopy- near-infrared (NIR), mid-infrared (MIR), or Raman- have also been widely adopted for diesel quality assessment. These techniques are known for their accuracy, speed, easiness of spectra acquisition, non-destructive nature, minimal sample quantity requirements, and direct determination with minimal or no need for sample preparation [17–27].

In particular, infrared spectroscopy (the NIR and/or MIR ranges) has become a powerful tool to tackle the complexity of diesel samples, offering valuable information about functional groups such as amine (-NH), hydroxyl (-OH), thiol (-SH), methylene (-CH_2), methyl (-CH_3), carbon-carbon (C-C), carbonyl associated groups (-C=O), carbon-oxygen (C-O), and aromatic (-C=C-). Due to its clear advantages, IR spectroscopy combined with multivariate calibration techniques, both linear and nonlinear, such as multiple linear regression (MLR), principal component regression (PCR), partial least squares regression (PLS), artificial neural networks (ANN), support vector machines (SVM) among others, are suitable approaches for obtaining quantitative information of diesel both off-line and on-line, as an alternative to standard methods. In that sense, the (relatively) low cost, the advances in instrumentation, and the progress in chemometric methods have boosted the use of NIR- and MIR-based analysis. Table I-4 shows several standard procedures for implementing these techniques in the quality control of petroleum products.

Table I-4. NIR and MIR as an ASTM-compliant tool. Adapted from [28].

Standard norm	Procedure
ASTM E1655: Standard Practice for Infrared Multivariate Quantitative Analysis.	Development of methods
ASTM D8321: Standard Practice for the Development and Validation of Multivariate Analyses for Use in Predicting Properties of Petroleum Products, Liquid Fuels, and Lubricants Based on Spectroscopic Detection.	Multivariate analysis of petroleum products
ASTM D6122: Standard Practice for Performance Validation of Inline, Inline, Field, and Laboratory Multivariate Infrared Spectrophotometers.	Method validation
ASTM D8340: Standard Practice for Performance-Based Qualification of Spectroscopic Analyzer Systems.	Validation of results

I.1.4 Application of PLS and ANN models based on IR to diesel fuels

Throughout the years, many efforts have been made to develop multivariate calibration models that allow the prediction of fuel properties from their IR spectrum. Among these models, PLS and ANN stand out as the most commonly applied linear and nonlinear calibration methods, respectively. The burgeoning interest in this field has resulted in many publications, making a comprehensive literature review daunting. Therefore, the literature review presented in this section focuses solely on the prediction of eleven properties in diesel and diesel/biodiesel samples using IR spectroscopy coupled with PLS and ANN over the last 25 years. The properties of interest are distillation temperatures at 65% (1), 85% (2), and 95% (3) recovered (T65%, T85%, and T95%), flash point (4), cloud point (5), density (6), cetane number (7), sulfur content (8), viscosity (9), cold filter plugging point (CFPP) (10), and fatty acid methyl esters (FAME) (11). Web of Science, Scopus, Google Scholar, and Microsoft Academic are the sources consulted.

In 1999, Fodor et al. [29] used mid-IR spectroscopy in attenuated total reflectance (ATR) mode combined with PLS regression to predict twenty-seven physicochemical properties, encompassing viscosity and density of middle distillates (diesel and kerosene). Several preprocessing methods were evaluated, and the ASTM-E-1655 [30] was used for the first time to verify the agreement between the PLS model predictions and the reference ASTM methods. Afterward, Soyemi et al. [31] achieved a better

predictive ability with NIR data using PCR compared to PLS. Subsequently, Santos Jr. et al. [32] showcased that PLS and ANN, in conjunction with FTIR-ATR, NIR, and Raman, could successfully predict various quality parameters. The ANN/FT-Raman models exhibited the best predictive ability among all the models they considered.

In 2006, Pasadakis et al. [33] predicted the cold properties and a distillation profile of diesel fuel fractions using MIR spectroscopy and ANN with an accuracy comparable to the respective standard method. Furthermore, Oliveira et al. [34] developed NIR/PLS and NIR/ANN models to quantify biodiesel content in biodiesel/diesel blends. Bezerra de Lira et al. [35] developed PLS based on NIR and MIR spectroscopy to predict the quality parameters of diesel/biodiesel blends. Similarly to previous studies [29,32], the root mean square error of prediction (RMSEP) values were comparable with the reproducibility of the standard method. In 2010, Gonzaga and Pasquini [36] used a low-cost short wave NIR spectrophotometer to predict the quality parameters of diesel fuel. Other authors, such as Nespeca et al. [37], have employed multivariate filters and variable selection techniques to obtain the best model.

Over the years, the literature has focused on estimating physicochemical properties of diesel/biodiesel blends, improving variable selection techniques to obtain the best predictive ability of the model incorporating other nonlinear techniques such as SVM [38–40], support vector regression (SVR) [10,20], and so on. Recently, data fusion has also gained attention to investigate whether complementing chemical information from different data sources could improve the estimation of properties in crude oil [41–43] and diesel [44–46]. IR or nuclear magnetic resonance (NMR) spectra have been mainly used in this sense. However, the inherent variability within the processes might influence the models' performance.

Table I-5 summarizes the studies that used PLS and ANN chronologically. For each property, we identified the references, date, the number of samples used, the type of samples, the regression method and the analytical techniques, the spectral region used, the preprocessing technique, the number of LVs or nodes for developing the models, the most outstanding results obtained only for the models of interest, the type of validation and finally, the calibration range. Several properties were analyzed in some works. Some of the most studied properties were density, cetane number, sulfur content, viscosity, FAME content, and flash point. Other properties such as T85%, cloud point, T95%, T65%, and CFPP were less studied.

Following Table I-5, PLS was the most applied regression model. In contrast, only five articles [32–34,47,48] applied ANN models with IR data to predict distillation temperatures, density, viscosity, cloud point, cetane number, FAME, and sulfur content

in diesel samples. Comparatively, studies based on the estimation of diesel properties using other diesel properties as inputs of ANN instead of IR spectra have received much more attention in the literature [48–60]. Also, regardless of the input data of the ANN, the studies for predicting biodiesel properties [61–69] are more common than the studies for diesel properties.

Table I-5 also indicates that the most used spectral pretreatments were the first derivative and mean centering, followed by vector normalization. Moreover, FTIR spectroscopy and the NIR region have been the most frequently used spectroscopic techniques and spectral regions. Among the studies listed, only two compared PLS and ANN model performance based on different spectroscopies. In one of them, Santos Jr. et al. [32] compared the performance of PLS and ANN models based on FTIR-ATR, FTNIR, and FT-Raman spectroscopies to predict properties such as density, T85% recovered, viscosity, and sulfur. They found that the PLS model based on the NIR region yielded better predictions for density, viscosity, and sulfur content with an RMSEP equal to 0.23 kg/m^3, 0.08 mm^2/s, and 0.012 %w/w, respectively. In contrast, for predicting T85% recovered, the FT-Raman/ANN models yielded better predictions with an RMSEP of 2.22 ºC. In the other study, Oliveira et al. [34] only compared the performance of PLS and ANN models based on FTIR-ATR and FTNIR spectroscopies to predict biodiesel content (FAME). They obtained a better prediction of biodiesel content with a PLS model in the NIR spectral region (RMSEP = 0.061 %v/v). In addition, the ANN model with NIR or MIR spectra had the same predictive ability for this property. The results of both works suggested that, for some properties, the PLS model might exhibit better predictive ability for NIR data.

Bezerra de Lira et al. [35] and Pimentel et al. [70] also compared the performance of PLS models based on MIR and NIR spectra for predicting sulfur content, density, T85%, and biodiesel content. According to Bezerra de Lira et al. [35], the NIR region was the more suitable to predict density and T85%, with RMSEP of 0.56 kg/m^3 and 2.01 ºC, respectively. In contrast, the MIR region was the best for predicting the sulfur content with PLS (RMSEP of 0.01 %w/w). Pimentel et al. [70] achieved a better prediction with NIR data and the PLS model for predicting the biodiesel content (RMSEP of 0.18 %v/v).

The literature review showed that if the property estimated is highly related to IR spectra, the predictions from PLS models based on IR spectra have good accuracy and perform better than or similar to ANN models. For other properties less related to IR spectra, ANN performed better than PLS models. Furthermore, the number of samples employed in the studies ranged from 50 to 684. Notably, many studies used fewer than

200 samples, highlighting the difficulty in obtaining samples that spanned the full range of compositional variability expected in production and finished product fuels.

The property ranges corresponding to the studies that reported the best results were 846.5-872.2 kg/m^3 (density), 300-360 ºC (T85%), 47.1-49.7 (cetane number), 2.2-3.8 mm^2/s (viscosity), -12.4-2.2 ºC (cloud point), -47-6 ºC (CFPP), 300-2100 mg/kg (sulfur content) and 0.2-30 %v/v (FAME content). The lowest RMSEP values for density, T85%, T95%, cetane number, viscosity, cloud point, CFPP, sulfur content and FAME content were 0.23 kg/m^3 [32], 2.01 ºC [35], 4.03 ºC [71], 0.167 [72], 0.074 mm^2/s [73], 1.132 ºC [72], 1.15 ºC [71], 2.17 ºC [22], 0.01 mg/kg [35] and 0.16 %(v/v) [74], respectively. The standard prediction error (SEP) referenced for T65% was 0.9 ºC (280–356ºC) [33]. However, aside from the cetane number and cloud point, these RMSEP values were calculated from external validation sets containing at most 45 samples.

As can be seen in Table I-5 and the databases consulted, there is no literature for properties such as flash point and CFPP describing their estimation by ANN models and IR spectroscopy. Furthermore, only in Ref. [33] was an ANN model based on MIR spectroscopy used to predict one (T65%) of the reviewed distillation temperatures. Regarding the cold properties, the available literature is especially scarce. In this sense, the prediction of CFPP in diesel samples, for instance, was solely reported in Ref. [22], employing the more prevalent approach based on PLS and NIR data.

Another noteworthy observation from Table I-5 is the growing number of studies on diesel samples blended with biodiesel over the years. These blends represent a viable option to expand the use of renewable energy sources. Biodiesel is a renewable energy source [75] and is considered the main substitute for diesel because it is free of sulfur and aromatic compounds and has a high cetane number [76]. In this context, there has been a marked interest in determining the quality parameters of biodiesel samples employing multivariate models with IR [77], with the CFPP being one of the most studied [70,78–80].

While the research above showcased the suitability of PLS/IR and ANN/IR methods for predicting diesel properties, it is crucial to acknowledge that their implementation in the industry may not be straightforward. Routine implementation requires that prediction errors be within the limits admitted by the reference method and also have tools that allow overcoming the loss of predictive ability of the model when working conditions vary. Due to this, implementing IR spectroscopy for routine use in diesel quality monitoring operations is challenging.

Table I-5. Review of PLS and ANN methods based on IR spectra used to determine 11 physicochemical diesel properties.

Ref.	Year	Samples (calibration, validation)		Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Density (kg/m^3)												
Fodor et al. [29]	1999	684 (547, 137)		commercial diesel	PLS	FTIR	4000-650	MC	19	SEP_cv = 0.9 R^2cv = 0.9952	CV	788.0-871.9
Soyemi et al. [31]	2000	118		diesel	PLS	NIR	13330-4000	MC	2	RMSEP = 6 R^2 = 0.985	LOOCV	800-870
Santos Jr. et al. [32]	2005	90	(45,45)	production diesel	PLS	FTIR-ATR1[a]	660–1490 1560–1660 2940–3100[a]	VN	10	RMSEP = 0.45[a]	external validation	846.5-872.2
						FTIR-ATR2[b]	1350–1730 2750–3090[b]	1st derivative	3	RMSEP = 2.2[b]		
						FTNIR[c]	4440–4760 5230–6030 8130–8430[c]	VN	10	RMSEP = 0.23[c]		
						FT-Raman[d]	1220–1520 2620–3060[d]		6	RMSEP = 1.2[d]		
			(60,30)		ANN	FTIR-ATR1[a]*	660–1490 1560–1660 2940–3100[a]		32-32-1	RMSEP = 0.96[a]*	CV	
						FTNIR[c]*	4440–4760 5230–6030 8130–8430[c]		28-28-1	RMSEP = 0.43[c]*		
						FT-Raman[d]*	1220–1520 2620–3060[d]		31-31-1	RMSEP = 1.2[d]*		
Morris et al. [81]	2009	280		synthethic diesel	PLS	NIR	10000-6250	baseline + MC	8	r^2 = 0.96 RMSECV = 2.4	LOOCV	740-880

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Bezerra de Lira et al. [35]	2010	161 (86,30)	synthethic blend [1] biodiesel [2] biodiesel diesel	PLS	FTIR[1]	12 000 –4000[a] 4000–600[b]	SG1D (a second-order polynomial and an 11-point window)	7[a] 8[b]	[a]RMSEP = 0.60[1] [b]RMSEP = 1.58[1]	full CV	836–860
					FTIR-ATR[2]			8[a] 8[b]	[a] RMSEP = 0.56[2] [b]RMSEP = 1.60[2]		
					FTIR portable, 2	1789.4– 650[c]		10[c]	[c]RMSEP =1.45[2]		
Ferrão et al. [74]	2010	85 (57, 28)	biodiesel/ diesel blends	PLS	HATR-FTIR	1070-650	MSC + MC	7	RMSECV = 0.54 RMSEP = 0.54	LOOCV, external validation	848.2–866.2
Marinović et al. [73]	2012	93	production diesel	PLS	FTIR-ATR	3500–2500 1670–650 [a]	MC	13[a] 9[b]	R^2 = 0.980[a] RMSECV = 0.567[a] R^2 = 0.960[b] RMSECV = 0.836[b]	LOOCV	827.5 – 840
					FT-Raman*	1670–650 [b]		6[a] 9[b]	R^2 = 0.952[a]* RMSECV = 0.892[a]* R^2 = 0.934[b]* RMSECV = 1.07[b]*		
Bolanca et al. [48]	2012	93	commercial diesel	ANN	FTIR-ATR	4000–650	-	8 hidden layer neurons	Abs. error mean = 0.7467 R = 0.9315	external validation	827.2–841.3
					FT- Raman	3700–350		8 hidden layer neurons	Abs. error mean = 0.9399 R = 0.9092		
Brouillette et al. [82]	2016	166 (88%, 12%)	production diesel	PLS	FT-NIR	9100-6500	MSC + first derivative	6	RMSECV = 2.25 R^2 CV= 0.95	venetian blinds CV	820–880

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Palou et al. [71]	2017	278	Biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	SG2D + SNV	4	RMSEC = 1.2 RMSEV = 1.6 RMSEP = 1.6	CV	826.2-861.7
Nespeca et al. [37]	2018	409 (271, 133)	production diesel	PLS	ATR-FTIR	3100-2450 1950-500	multivariate filter OSC	9	RMSECV = 0.5 RMSEP = 0.4 R^2 CV = 0.993 R^2 val = 0.994	venetian blinds CV	831.6-860.4
Al-kaf et al. [47]	2018	243 (122,121)	diesel	ANN	NIR	13330-6450	-	-	RMSEP = 0.828	-	781.8-872.8
Mishra et al. [44]	2021	395 (237, 158)	diesel	PLS	NIR	13330-6450	SNV and 2nd derivative	5	RMSEP = 2 R^2p = 0.96	10-fold CV	790-870
Distillation temperatures at 65%, 85% and 95% recovered (ºC)											
Fodor et al. [29]	1999	684 (547, 137)	commercial diesel	PLS	FTIR	4000-650	MC	16	SEP_cv = 8 R^2cv = 0.7932	CV	T95% 230-375
Santos Jr. et al. [32]	2005	90 (45,45)	production diesel	PLS	FTIR-ATR1[a]	660–900 1100–1770[a]	VN	10	RMSEP = 2.9[a]	external validation	T85% 245.6-369.8
					FTIR-ATR2[b]	1070-1590 2820–2990[b]	1st derivative	2	RMSEP = 4.5[b]		
					FTNIR[c]	4440–4760 5230–6030 8130–8430[c]	VN	8	RMSEP = 2.9[c]		
					FT-Raman[d]	1220–1660 2790–2990[d]		6	RMSEP = 3.7[d]		

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Santos Jr. et al. [32]	2005	90 (60,30)	production diesel	ANN	FTIR-ATR1[a]*	660–1490 1560–1660 2940–3100[a]	VN	32-32-1	RMSEP = 3.41[a]*	CV	T85% 245.6-369.8
					FTNIR[c]*	4440–4760 5230–6030 8130–8430[c]		28-28-1	RMSEP = 2.9[c]*		
					FT-Raman[d]*	1220–1520 2620–3060[d]		31-31-1	RMSEP = 2.22[d]*		
Pasadakis et al. [33]	2006	122 (85%, 15%)	production diesel	ANN	FTIR	1474–1464 1314–1306 810–800 740–720	scale	28-2-1-1	SEPtraining = 1.3[a] SEPtest = 0.9[a]	-	[a]T65% 280–356
									SEPtraining = 1.2[b] SEPtest = 1.0[b]		[b]T85% 314–377
									SEPtraining = 1.0[c] SEPtest = 1.3[c]		[c]T95 347–401
Bezerra de Lira et al. [35]	2010	161 (86,30)	synthethic blend [1] biodiesel [2] biodiesel+ diesel	PLS	FTIR[1]	12 000 –4000[a] 4000–600[b]	SG1D (a second-order polynomial and an 11-point window)	8[a] 8[b]	[a]RMSEP = 2.01[1] [b]RMSEP = 3.28[1]	full CV	T85% 300–360 °C
					FTIR-ATR[2]			3[a] 5[b]	[a] RMSEP = 3.37[2] [b]RMSEP = 4.38[2]		
					FTIR [portable, 2]	1789.4– 650[c]		9[c]	[c]RMSEP =4.18[2]		
Gonzaga and Pasquini [36]	2010	93 (60, 30)	diesel	PLS	SW-NIR	11630–9700	SG1D (a second order polynomial and 21 point window) + MC.	6	RMSECV = 3.9 RMSEP = 3.2 rcv = 0.85 rp = 0.90	full CV	T85% 329.5-363.2

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Palou et al. [71]	2017	278	Biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	SG2D + SNV	8	RMSEC = 1.18 RMSEV = 5.39 RMSEP = 4.03		T95% 336.5-384.9
Nespeca et al. [37]	2018	409 (271, 133)	production diesel	PLS	ATR-FTIR	3100-2450 1950-500	multivariate filter GLSW	8	RMSECV = 5.1 RMSEP = 4.7 R^2 CV = 0.593 R^2 val = 0.639	venetian blinds CV	T85% 330-369.5
Cetane Number											
Fodor et al. [29]	1999	684 (547, 137)	commercial diesel	PLS	FTIR	4000-650	MC	12	SEP_cv = 1.6 R^2cv = 0.7641	CV	36.9-61.3
Soyemi et al. [31]	2000	118	diesel	PLS	NIR	13330-4000	MC	3	RMSEP = 1.23 R^2 = 0.985	LOOCV	40-60
Alves et al. [10]	2012	114 (77 and 37)	diesel	PLS	MID/NIR	3500–4678	baseline + MC	5	RMSEC = 0.745 RMSEP = 0.556 R^2 = 0.860		37.6–48.9
Marinović et al. [73]	2012	93	production diesel	PLS	FTIR-ATR	3500–2500 1670–650 [a]	MC	11[a] 9[b]	R^2 = 0.982[a] RMSECV = 0.279[a] R^2 = 0.984[b] RMSECV = 0.267[b]	LOOCV	50-56
					FT-Raman*	1670–650 [b]		7[a] 6[b]	R^2 = 0.980[a]* RMSECV = 0.316[a]* R^2 = 0.972[b]* RMSECV = 0.362[b]*		

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Bolanca et al. [48]	2012	93	commercial diesel	ANN	FTIR-ATR	4000–650	-	8 hidden layer neurons	Abs. error mean = 0.3228 R = 0.9599	external validation	50.1–55.9
					FTIR-Raman	3700–350			Abs. error mean = 0.3669 R = 0.9406		
Brouillette et al. [82]	2016	166 (88%, 12%)	production diesel	PLS	FT-NIR	9000-6500	1st derivative + an 11-point, third-degree polynomial (SG) + MSC	5	RMSECV = 1.1 R^2 CV= 0.91	venetian blinds CV	42 - 53
Zhan and Yang [38]	2017	381 (254,127)	diesel	PLS	NIR	13330-6450	none	-	RMSEC = 1.85 RMSEP = 2.14	external validation	20.4 49.5
Palou et al. [71]	2017	278	biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	SG2D + SNV	2	RMSEC = 0.45 RMSEV = 1.02 RMSEP = 0.7	-	45.5 - 59
Al-kaf et al. [47]	2018	232 (116,116)	diesel	ANN	NIR	13330-6450	-	-	RMSEP = 1.87	-	40.3-61.3
Nespeca et al. [37]	2018	409 (271, 133)	production diesel	PLS	ATR-FTIR	3100-2450 1950-500.	multivariate filter OSC	7	RMSECV = 0.6 RMSEP = 0.6 R^2 CV = 0.815 R^2 val = 0.841	venetian blinds CV	42.2-51
Barra et al. [25]	2020	50 (40, 10)	production diesel	PLS	FTIR	4000-400	baseline correction + MC and SG1D	8	RMSEC = 0.28 RMSEP = 0.42 R^2 = 0.99	LOOCV	49-59

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)		Type of samples	Regression Method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Msimanga et al. [72]	2022	80 (56, 24)		commercial diesel	PLS	FTIR-ATR	4000–650	baseline correction + normalization	-	RMSEP = 0.167 R^2 = 0.965	CV	47.1-49.7
Viscosity (mm^2/s)												
Fodor et al. [29]	1999	684 (547, 137)		diesel commercial	PLS	FTIR	4000-650	MC	19	SEP_cv = 0.08 R^2cv = 0.9640	CV	1.14-4.05
Soyemi et al. [31]	2000	118		diesel	PLS	NIR	13330-4000	MC	2	RMSEP = 0.23 R^2 = 0.979	LOOCV	1.3-3.4
Santos Jr. et al. [32]	2005	90	(45,45)	production diesel	PLS	FTIR-ATR1[a]	660–1660[a]	1st derivative + VN	5	RMSEP = 0.09[a]	external validation	2.6-5.3
						FTIR-ATR2[b]	1200–1590 2690–3090[b]	no pre-processing	5	RMSEP = 0.2[b]		
						FTNIR[c]	4440–4760 5230–6030 8130–8430[c]	vector normalization	8	RMSEP = 0.08[c]		
						FT-Raman[d]	970-1120 1220–1520 2620–3060[d]	no pre-processing	6	RMSEP = 0.16[d]		
			(60,30)		ANN	FTIR-ATR1[a]*	660–1660[a]	1st derivative + VN	32-32-1	RMSEP = 0.23[a]*	CV	
						FTNIR[c]*	4440–4760 5230–6030 8130–8430[c]	VN	28-28-1	RMSEP = 0.16[c]*		
						FT-Raman[d]*	970-1120 1220–1520 2620–3060[d]	no pre-processing	31-31-1	RMSEP = 0.15[d]*		

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Morris et al. [81]	2009	261	synthethic diesel	PLS	NIR	10000-6250	baseline + MC	8	r^2 = 0.85 RMSECV = 0.195	LOOCV	-
Marinović et al. [73]	2012	93	production diesel	PLS	FTIR-ATR	3500–2500 1670–650 [a]	MC	12[a] 7[b]	R^2 = 0.989[a] RMSECV = 0.074[a] R^2 = 0.983[b] RMSECV = 1.07[b]*	LOOCV	2.2-3.8
					FT-Raman*	1670–650 [b]		8[a] 5[b]	R^2 = 0.972[a]* RMSECV = 0.116[a]* R^2 = 0.963[b]* RMSECV = 0.133[b]*		
Bolanca et al. [48]	2012	93	commercial diesel	ANN	FTIR-ATR	4000–650	-	8 hidden layer neurons	Abs. error mean = 0.7467 R = 0.9315	external validation	2.24-3.79
					FT- Raman	3700–350			Abs. error mean = 0.9399 R = 0.9092		
Brouillette et al. [82]	2016	134 (88%, 12%)	production diesel	PLS	FT-NIR	9100-6500	MSC + 1st derivative	6	RMSECV = 0.17 R^2 CV= 0.85	venetian blinds CV	2-4.5
Al-kaf et al. [47]	2018	232 (116,116)	diesel	ANN	NIR	13330-6450	-	-	RMSEP = 0.09	-	1.3-3.55
Mishra et al. [44]	2021	395 (237 and 158)	diesel	PLS	NIR	13330-6450	SNV and 2nd derivative	5	RMSEP = 0.21 R^2p = 0.78	10-fold CV	1-4
Hradecká et al. [22]	2021	90	diesel	PLS	FTIR/NIR	9000–5000		10	RMSEC = 0.14 R^2 = 0.992	10-fold CV	1.13-4.88

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Flash Point (ºC)											
Fodor et al. [29]	1999	684 (547, 137)	commercial diesel	PLS	FTIR	4000-650	MC	9	SEP_cv = 5 R^2cv = 0.5203	CV	23-96
Morris et al. [81]	2009	280	synthethic diesel	PLS	NIR	10000-6250	baseline + MC	4	r^2 = 0.22 RMSECV = 8.9	LOOCV	-
Ferrão et al. [74]	2010	85 (57, 28)	biodiesel/ diesel blends	PLS	HATR-FTIR	968-756	MSC + MC	8	RMSECV = 0.93 RMSEP = 0.90	LOOCV and external validation	47.0-79.5
Alves et al. [10]	2012	451 (350, 101)	diesel	PLS	MID/NIR	3944–4769	SNV	3	RMSEC = 4.21 RMSEP = 3.77 R^2 = 0.698	CV	24.5-76.5
Brouillette et al. [82]	2016	107 (88%, 12%)	production diesel	PLS	FT-NIR	9100-6500	MSC + 1st derivative	3	RMSECV = 6.7 R^2 CV= 0.37	venetian blinds CV	65-95
Palou et al. [71]	2017	278	biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	Savitzky–Golay's second derivative	7	RMSEC = 0.57 RMSEV = 3.72 RMSEP = 2.90	CV	56-71
Nespeca et al. [37]	2018	409 (271, 133)	production diesel	PLS	ATR-FTIR	3100-2450 1950-500	multivariate filter OSC	7	RMSECV = 2 RMSEP = 2 R^2 CV = 0.809 R^2 val = 0.791	venetian blinds CV	8-70
Msimanga et al. [72]	2022	80 (56, 24)	commercial diesel	PLS	FTIR-ATR	4000–650	baseline correction + normalization	-	RMSEP = 1.132 R^2 = 0.956	CV	-

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Cloud Point (ºC)											
Fodor et al. [29]	1999	684 (547, 137)	commercial diesel	PLS	FTIR	4000-650	MC	16	SEP_cv = 3.1 R^2cv = 0.8430	CV	-60.5 – 2.1
Pasadakis et al. [33]	2006	122 (85%, 15%)	production diesel	ANN	FTIR	1700–600	1st derivative	6-5-3-1	SEPtraining = 1.4 SEPtest = 3.1	-	-9-17
Brouillette et al. [82]	2016	111 (88%, 12%)	production diesel	PLS	FT-NIR	9100-6500	MSC + 1st derivative	5	RMSECV = 3.6 R^2 CV= 0.68	venetian blinds CV	-25-15
Palou et al. [71]	2017	278	biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	SG2D	7	RMSEC = 0.53 RMSEV = 1.22 RMSEP = 1.15	CV	-12.4 - 2.2
CFPP (ºC)											
Hradecká et al. [22]	2021	90	diesel	PLS	FTIR/NIR	9000–5000	-	10	RMSEC = 2.17 R^2 = 0.988	10-fold CV	−47 - 6 ºC
Sulfur Content											
Breikreitz et al. [83]	2003	97 (41,30,26)	diesel	PLS	NIR	12500-5560	SG1D (a second-order polynomial)	6	RMSEP = 360	internal and external validation	700-3300
Santos Jr. et al. [32]	2005	90 (45,45)	production diesel	PLS	FTIR-ATR1[a]	660–1490[a]	1st derivative + VN	8	RMSEP = 160[a]	external validation	1900-3500
					FTIR-ATR2[b]	1200–1660 2430–3090[b]		1	RMSEP = 220[b]		
					FTNIR[c],	4440–4760 5230–6030 8130–8430[c]		8	RMSEP = 120[c]		

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)		Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Santos Jr. et al. [32]	2005	90	(45,45)	production diesel	PLS	FT-Raman[d]	670-790 1220–1660 2790–3060[d]	VN	4	RMSEP = 150[d]	external validation	1900-3500
			(60,30)		ANN	FTIR-ATR1[a]*	660–1490[a]	1st derivative + VN	32-32-1	RMSEP = 100[a]*	CV	
						FTNIR[c]*	4440–4760 5230–6030 8130–8430[c]		28-28-1	RMSEP = 300[c]*		
						FT-Raman[d]*	670-790 1220–1660 2790–3060[d]	VN	31-31-1	RMSEP = 100[d]*		
Bezerra de Lira et al. [35]	2010	161 (86,30)		synthethic blend [1] biodiesel blends. [2] biodiesel blends + diesel	PLS	FTIR[1]	12 000 –4000[a] 4000–600[b]	SG1D (a second-order polynomial and an 11-point window)	6[a] 8[b]	[a]RMSEP = 200[1] [b]RMSEP = 200[1]	full CV	300–2100
						FTIR-ATR[2]			2[a] 2[b]	[a] RMSEP = 200[2] [b]RMSEP = 100[2]		
						FTIR portable, [2]	1789.4– 650[c]		10[c]	[c]RMSEP = 200[2]		
Soares et al. [84]	2010	-		production diesel	PLS	ATR-FTIR	2717-1145 856-665	-	8	RMSEP = 0.038	-	400-2500
							3099-2713			RMSEP = 0.16		
Ferrão et al. [74]	2010	85 (57, 28)		Biodiesel/ diesel blends	PLS	HATR-FTIR	1491-1070 3165-2746	MSC + MC	9	RMSECV = 13.4 RMSEP = 13.9	LOOCV and external validation	312-1351

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Palou et al. [71]	2017	278	biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	SG2D + SNV	5	RMSEC = 0.48 RMSEV = 2.56 RMSEP = 1.88	CV	2-22
Nespeca et al. [37]	2018	409 (271, 133)	production diesel	PLS	ATR-FTIR	3100-2450 1950-500	multivariate filter OSC	5	RMSECV = 100 RMSEP = 100 R^2 CV = 0.969 R^2 val = 0.974	venetian blinds CV	100 - 1900
Correia et al. [85]	2018	133 (89, 44)	diesel blends	PLS	MicroNIR	11000-6000	SG1D (a second-order polynomial and an 7-point window)	5	RMSEP = 13.2 R^2 = 0.992	venetian blinds 5-fold CV	10– 500
Hradecká et al. [22]	2021	90	diesel	PLS	FTIR/NIR	10000–6000	-	10	RMSEC = 0.53 R^2 = 0.984	10-fold CV	0.36-11
Msimanga et al. [72]	2022	80 (56, 24)	commercial diesel	PLS	FTIR-ATR	4000–650	baseline correction + normalization	-	RMSEP = 0.26 R^2 = 0.979	CV	5.0-11.7
Zheng et al. [86]	2023	95 (70%,30%)	synthethic diesel	PLS	FTIR	12 490- 3600	SG1D (15-point, second order)		RMSEC = 47.6 R^2c = 0.97 RMSEP = 50.7 R^2 = 0.97	-	10.3–1038.0
FAME Content % (v/v)											
Pimentel et al [70]	2006	(43, 13)	biodiesel/diesel blends	PLS	MID	4500-500	1st derivative	3	RMSEP = 0.25	external validation	0-5 %(v/v)
					NIR	13330-4000		6	RMSEP = 0.18		

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression Method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Oliveira et al. [34]	2006	96 (48, 48)	diesel/ biodiesel blends	PLS	FTIR-ATR[a] FTNIR[b]	3699-2974 1841-925[a] 8930-7927, 7479-6685, 6137-4463[b]	1st derivative	10	RMSEP = 0.202[a] RMSEP = 0.061[b]	external validation	0–100% (w/w)
				ANN				-	RMSEP = 0.213[a] RMSEP = 0.212[b]		
Ferrão et al. [74]	2010	85 (57, 28)	biodiesel/ diesel blends	PLS	HATR-FTIR	1909-650 3165-2746	MSC + MC	5	RMSECV = 0.21 RMSEP = 0.16	LOOCV and external validation	0.2-30 %(v/v)
Alves et al. [87]	2013	91	diesel blends	PLS	NIR	5500-6000	SNV	4	RMSEP% = 0.25 R^2 = 0.9999	CV	0-20 %(v/v)
Alves et al. [39]	2016	101 (66, 35)	diesel blends of production	PLS	MID/NIR	3900–9000	SNV	4	RMSEC % = 0.17 RMSEP % = 0.24 R^2 = 0.999		0–25.5 %(v/v)
Palou et al. [71]	2017	278	biodiesel /diesel blends of production	PLS	FT-NIR	10000-4550	SG2D	4	RMSEC = 0.63 RMSEV = 0.67 RMSEP = 0.51	CV	0-14.9 %(v/v)
Nespeca et al. [37]	2018	409 (271, 133)	production diesel	PLS	ATR-FTIR	3100-2450 1950-500	multivariate filter OSC	6	RMSECV = 0.2 RMSEP = 0.2 R^2 CV = 0.877 R^2 val = 0.878	venetian blinds CV	0.6-7.4 %(v/v)

Table I-5 (cont.)

Ref.	Year	Samples (cal., val.)	Type of samples	Regression method	Spectroscopic technique	Spectral region (cm^{-1})	Pretreatment	LVs or nodes	Error	Validation	Calibration range
Correia et al. [85]	2018	133 (89, 44)	biodiesel/ diesel blends	PLS	MicroNIR	11000-6000	SG1D (a second-order polynomial and an 7-point window)	5	RMSEP = 1.8 wt% R^2 = 0.998	venetian blinds 5-fold CV	0-100 wt%

PLS: Partial Least Squares
ANN: Artificial Neural Network
ATR-FTIR: Attenuated Total Reflectance-Fourier Transform Infrared
FTIR: Fourier transform infrared
HATR: horizontal attenuated total reflectance
MC: Mean centering
MSC: Multiplicative scatter correction
VN: Vector normalization
SG1D: Savitzky-Golay's first derivative
OSC: orthogonal signal correction
SG2D: Savitzky-Golay's second derivative
SNV: standard normal variate
GLSW: Generalized Least Squares weighting
K-ANN: Kohonen artificial neural network
LVs: latent variables
R_c^2: coefficient of determination of calibration
R_{cv}^2: coefficient of determination of cross-validation
R_p^2: coefficient of determination of prediction
RMSEC: root mean square error of calibration
RMSECV: root mean square error of cross-validation
RMSEP: root mean square error of prediction
SPE: Standard Prediction Error
r: correlation coefficient
LOOCV: leave-one-out cross-validation
CV: cross-validation
EV: External validation

I.1.5 Transfer of multivariate calibration models in diesel analysis

Several methodologies have been developed to extend the generalizability of predictive models when applied to new spectra acquired under different environmental conditions or with different spectrophotometers [88–90]. Among these, spectral transformation or combinations of transforms, including derivatives, standard normal variate (SNV), multiplicative scatter correction, and filtering algorithms, aim to correct baseline shifts and changes in signal-to-noise ratio. Besides, calibration transfer methods, such as spectral transfer and model transfer, are more commonly employed to remove changes in an instrument's bandwidth or dispersion and deviations in a detector's response linearity. For spectral transfer, piecewise direct standardization (PDS), direct standardization (DS), and orthogonal signal correction (OSC), among other variants, adjust the new spectra to resemble the spectra measured on the instrument used for initial calibration. As an alternative to spectral transfer, model transfer involves adapting the existing calibration model for a new spectrophotometer, ranging from model updating (MU) to slope-bias correction and transfer of the regression equation.

The transfer of PLS models utilizing IR data for fuel analysis has been previously explored in the literature [91]. Specifically in diesel analysis, Perston and Harris [92] assessed the suitability of calibration-transfer approaches (bias correction and model updating) between instruments of a PLS model used to determine the biodiesel content in diesel/biodiesel blends by ASTM D7371-14. The performances of calibration transfer approaches were assessed through external validation using F-tests. In addition, Bezerra de Lira et al. [35] used DS to address the instrument calibration transfer of PLS models for predicting properties such as biodiesel content and sulfur content, density, and T85% recovered of diesel/biodiesel blends. In their study, ten transfer samples were used, and the prediction errors obtained after applying DS between two spectrophotometers were similar to those obtained by full recalibration of the secondary spectrophotometer.

Cooper et al. [93] proposed a novel calibration transfer approach utilizing virtual standards and the slope and bias correction (SBCP) method. This approach aimed to overcome the loss of the predictive ability of PLS models when predicting ten diesel properties in a secondary spectrophotometer. The virtual standards were mathematically created from seven spectra of pure solvents acquired on primary and secondary spectrophotometers to remove the need to keep fuel standards or produce complex mixtures to replicate fuels. Furthermore, the predictive ability of the PLS models in the secondary spectrophotometer improved significantly after SBCP correction. Also, this improvement surpassed that obtained after PDS correction using transfer spectra of fuel.

Da Silva et al. [94] also used virtual standards as transfer samples to apply the reverse standardization (RS) method from a benchtop to a handheld NIR spectrometer in order to determine the biodiesel content of diesel/biodiesel blends. The adoption of virtual standards prevented alterations in the composition or volatilization during the transportation and storage of transfer fuel samples. However, the RMSEP values obtained after RS were not equivalent to the reproducibility of the reference method.

Table I-6 summarizes the above studies and shows the techniques used, calibration transfer scenarios, and properties examined. The most studied scenario was calibration transfer between instruments, and the most analyzed property was the biodiesel content. Two of the studies in this table adopted strategies requiring transfer samples measured on both primary and secondary instruments. In contrast, the other two studies adopted strategies based on virtual standards. Above all, the literature review revealed the scarcity of studies concerning the application of calibration transfer methods for PLS models in diesel analysis.

Table I-6. A summary of recent calibration transfer techniques applied to diesel and diesel/biodiesel blends.

Ref.	Year	Technique	Calibration transfer scenario	Properties
Perston and Harris [92]	2009	Bias correction and model updating	Between FTIR spectrometer with different ATR accessory	biodiesel content
Bezerra de Lira et al. [35]	2010	Direct standardization	Between two FTIR instruments	biodiesel content, density, sulfur content T85%
Cooper et al. [93]	2011	Virtual standard slope-bias transfer, PDS, slope-bias corrected	Between two NIR handheld fuel analyzers	API gravity, % aromatics, cetane index, T10%, T20%, T50%, T90%, flash point, % hydrogen, % saturates
Da Silva et al. [94]	2017	Virtual standards as transfer samples in the reverse standardization (RS) method	Between a high-resolution benchtop FT-NIR and a handheld MicroNIR	biodiesel content

I.1.6 Motivation

Previous studies about diesel analysis reveal that coupling IR spectroscopy with chemometric methods is a powerful approach for predicting physicochemical properties in diesel samples. The PLS model and its variants stand out as the most extensively employed regression technique among the chemometric methods. However, it is also acknowledged that non-linear spectrum-property relationships in systems with strong intermolecular or intramolecular interactions, such as diesel fuel, could pose a challenge to the performance of the PLS model [95]. This is particularly relevant for predicting diesel properties that are not directly related to the chemical composition (e.g., T95%, flash point, cold filter plugging point) and, thus, to the spectrum [96].

Unlike the PLS model, nonlinear methods are deemed more suitable for building robust calibration models in complex systems [97,98]. In this sense, ANNs are among the most effective and popular nonlinear methods. ANNs possess the capability to approximate any linear or nonlinear relationship between input and output values by appropriately adjusting the free parameters or weights [99,100]. Accordingly, this capability of ANN could be tentatively harnessed to develop calibration models and obtain better predictions of diesel properties.

Regardless of the calibration model used, the limits of the applicability domain (AD) must be well-defined to ensure that the model will generate reliable predictions for new samples. In the PLS calibration model, the applicability limits are commonly based on Hotelling's T^2 and Q statistics [101,102]. Other limits based on criteria, such as ASTM's RMSSR (Root Mean Square Spectral Residuals) and NND (Nearest Neighbor Distance) [30], have also been used to flag spectra outside the established limits as discordant spectra. Although measures based on the similarity among spectra apply to all types of models, those that consider the specific form of the model, such as Hotelling's T^2 and Q statistics, are preferred since they are related to how the spectrum is being used by the model. A similar system for defining the limits of applicability of multivariate regression based on ANNs has not been reported yet.

The routine use of multivariate calibration models in diesel analysis requires substantial effort not only to develop the calibration model but also to maintain it to ensure its performance over time or when the work conditions vary. In this context, calibration-model adaptations have been proposed to address this issue. Specifically, in fuel analysis, to avoid contamination problems during use or changes in composition due to the volatility of some components, several studies use virtual standards (digitally created from spectra of pure components) for calibration transfer of infrared-PLS models of diesel

blends. In this regard, domain invariant-PLS (di-PLS) and dynamic orthogonal projection (DOP) standard-free calibration transfer approaches stand out for their versatility and free availability and are promising for enhancing the transferability of PLS calibrations between instruments in diesel analysis. To the best of the author's knowledge, the application of both calibration transfer approaches in the prediction of quality parameters of diesel samples for correction of external influences associated with different instruments has not been reported in the literature.

I.1.7 Hypotheses

The hypotheses of the work developed in this doctoral thesis are the following:

- The chemical information in the IR region makes it possible to determine the physiochemical properties that characterize the quality of a diesel sample from its IR spectrum.
- Multivariate calibration models such as PLS and ANN provide the means to model the complex relationships between the IR spectrum and the diesel properties of interest.
- A broad set of diesel samples representative of several production months could contribute to improving the performance of predictive models based on IR spectroscopy.
- Modeling the nonlinear IR spectrum–property relationships by using ANNs can improve the poor predictive ability of PLS models for some diesel properties such as T95% recovered, flash point, and CFPP.
- Similar to the applicability limits of the PLS model based on the leverage and the spectral residuals, those of a feed-forward neural network (FFNN) model can be defined by the squared Mahalanobis distance and Q-residuals as a measure of the leverage and spectral residuals.
- Monitoring the performance of calibration models enables the revealing of any spectral variability that has not been accounted for.
- Calibration-model adaptations can be used to overcome the loss of PLS model performance when applied to spectra obtained from a spectrophotometer different from the one used in model development.

I.2 Objectives

The general objective of this doctoral thesis is to develop and validate multivariate calibration models for the determination of the quality properties of diesel from its IR spectrum.

To fulfill the main objective, the following specific objectives are proposed:

1. To characterize the samples in terms of their physicochemical properties and spectral features.
2. To develop a predictive model based on IR spectroscopy coupled with an ANN multivariate calibration model and establish a methodology to define its limits of the applicability domain (AD).
3. To develop PLS and ANN multivariate calibration models to predict eleven physicochemical properties in diesel samples, providing an alternative to conventional methods.
4. To compare the predictive ability of PLS and ANN models developed to estimate each of the properties of interest in the samples.
5. To develop a strategy that allows monitoring the performance of PLS models over time.
6. To implement a strategy to overcome the loss of the predictive ability of a PLS calibration model between instruments by applying novel approaches such as di-PLS, DOP, and MU.

I.3 Thesis structure

This doctoral thesis is structured in seven chapters.

Chapter I. Introduction, objectives, and structure. This chapter is dedicated to the introduction, objectives, and structure of the study carried out in this doctoral thesis. In order to understand the importance and complexity of the subject under study, this chapter provides an overview of the relevance of diesel as a fuel and presents a simplified view of its production process. The characteristic chemical composition of diesel is discussed, along with the properties that characterize its quality and the suitability of infrared spectroscopy for determining these properties. These ideas support the hypotheses. It then provides a brief literature survey on the application of infrared spectroscopy to determine diesel properties. Also, the hypothesis, the general and specific objectives, and the thesis structure are presented.

Chapter II. Theoretical foundation. This chapter briefly describes the theoretical foundations of the instrumental and chemometric techniques used in this doctoral thesis.

Chapter III. Exploration of properties and spectra. This chapter is dedicated to characterizing desulfurized and commercial diesel samples in terms of their physicochemical properties and spectral features. The correlations between properties are studied. The instrumentation used in the spectral acquisition is also described, and classical spectral analysis is performed to identify the most representative regions of the characteristic functional groups of the diesel composition. The multivariate analysis of the spectra is carried out to detect natural groups of the samples.

Chapter IV. Calibration model based on artificial neural network for density prediction. Defining the limits of its applicability domain. This chapter describes a strategy for establishing a calibration model based on a feed-forward neural network (FFNN) to predict the density of diesel samples using mid-infrared spectra and its limits of the applicability domain (AD). Here, AD was defined by two limits: 1) the 0.99 quantile of the squared Mahalanobis distance calculated from the network activations of the training set and 2) the 0.99 quantile of the reconstruction error of the training spectra using either an autoencoder network or a decoder network. Furthermore, the performances of the decoder and autoencoder networks were compared.

Chapter V. PLS vs. ANN calibration models for determining diesel quality parameters. This chapter describes the establishment of calibration models based on PLS and FFNN for predicting six quality parameters of desulphurized diesel and eleven quality parameters of commercial diesel. The study includes the methodology used to 1) select the spectral region that provides the best results, 2) set the applicability limits and the tolerance limits admitted for each property, and 3) consider the validity of the models from an industrial point of view. Additionally, the predictive abilities of PLS and FFNN calibration models are compared.

Chapter VI. Monitoring and maintenance of PLS models. This chapter is dedicated to monitoring and maintaining the performance of PLS models over time. Additionally, the suitability of three calibration-model adaptations from an existing calibration PLS for determining density in routine analysis between two infrared spectrophotometers are compared.

Chapter VII. Conclusions and perspectives. This chapter presents the main conclusions of this doctoral thesis. Also, some recommendations for future work on the

implementation of IR spectroscopy-based analysis methods in the field of diesel analysis are proposed.

I.4 References

[1] FuelsEurope Statistical Report 2023, 2024. https://www.fuelseurope.eu/statistics.

[2] CORES, Corporación de Reservas Estratégicas de Productos Petrolíferos, Estadísticas. (2022). https://www.cores.es/es/estadisticas.

[3] F.G. Becker, M. Cleary, R.M. Team, H. Holtermann, D. The, N. Agenda, P. Science, S.K. Sk, R. Hinnebusch, R. Hinnebusch A, I. Rabinovich, Y. Olmert, D.Q.G.L.Q. Uld, W.K.H.U. Ri, V. Lq, W.K.H. Frxqwu, E. Zklfk, L. V Edvhg, R.Q. Wkh, F.G. Becker, N. Aboueldahab, R. Khalaf, L.R. De Elvira, T. Zintl, R. Hinnebusch, M. Karimi, S.M. Mousavi Shafaee, D. O 'driscoll, S. Watts, J. Kavanagh, B. Frederick, T. Norlen, A. O'Mahony, P. Voorhies, T. Szayna, N. Spalding, M.O. Jackson, M. Morelli, B. Satpathy, B. Muniapan, M. Dass, P. Katsamunska, Y. Pamuk, A. Stahn, E. Commission, T.E.D. Piccone, M.K. Annan, S. Djankov, M. Reynal-Querol, M. Couttenier, R. Soubeyran, P. Vym, E. Prague, World Bank, C. Bodea, N. Sambanis, A. Florea, A. Florea, M. Karimi, S.M. Mousavi Shafaee, N. Spalding, N. Sambanis, 2021, فاطمی. ح World Oil Outlook 2045, Vienna, Austria, 2021. https://woo.opec.org/.

[4] D. Havard, Oil and Gas Production Handbook - An introduction to oil and gas production, transport, refining and petrochemical industry, ABB Oil and Gas, Oslo, Norway, 2013.

[5] J.G. Speight, Petroleum Refinery Processes, in: Kirk-Othmer Encycl. Chem. Technol., 2018: pp. 1–24. https://doi.org/10.1002/0471238961.1805060919160509.a01.pub3.

[6] MathPro, An introduction to petroleum refining and the production of ultra low sulfur gasoline and diesel fuel, Bethesda, Maryland, 2011. https://theicct.org/publication/an-introduction-to-petroleum-refining-and-the-production-of-ultra-low-sulfur-gasoline-and-diesel-fuel/.

[7] J. Sánchez, Purificación de parafinas de petróleo por hidrogenación catalítica, Universidad Complutense de Madrid, 2003. http://biblioteca.ucm.es/tesis/qui/ucm-t26589.pdf.

[8] A. Palou Garcia, Desarrollo de nuevas metodologías espectrales para el control analítico de productos y procesos petroquímicos y farmacéuticos, Universitat

Autónoma de Barcelona, 2014. http://hdl.handle.net/10803/285412.

[9] J. Bacha, J. Freel, A. Gibbs, L. Gibbs, G. Hemighaus, K. Hoekman, J. Horn, M. Ingham, L. Jossens, D. Kohler, D. Lesnini, J. McGeehan, M. Nikanjam, E. Olsen, R. Organ, B. Scott, M. Sztenderowicz, A. Tiedemann, C. Walker, J. Lind, J. Jones, D. Scott, J. Mills, Diesel Fuels Technical Review, 2007. https://www.chevron.com/-/media/chevron/operations/documents/diesel-fuel-tech-review.pdf.

[10] J.C.L. Alves, C.B. Henriques, R.J. Poppi, Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system, Fuel. 97 (2012) 710–717. https://doi.org/10.1016/j.fuel.2012.03.016.

[11] P. Saxena, S. Jawale, M.H. Joshipura, A review on prediction of properties of biodiesel and blends of biodiesel, Procedia Eng. 51 (2013) 395–402. https://doi.org/10.1016/j.proeng.2013.01.055.

[12] S.K. Hoekman, A. Broch, C. Robbins, E. Ceniceros, M. Natarajan, Review of biodiesel composition , properties , and specifications, Renew. Sustain. Energy Rev. 16 (2012) 143–169. https://doi.org/10.1016/j.rser.2011.07.143.

[13] K.B. Murali, J.M. Mallikarjuna, Properties and performance of cotton seed oil-diesel blends as a fuel for compression ignition engines, J. Renew. Sustain. Energy. 1 (2009) 023106. https://doi.org/10.1063/1.3117342.

[14] M. Gülüm, A. Bilgin, Measurements and empirical correlations in predicting biodiesel-diesel blends' viscosity and density, Fuel. 199 (2017) 567–577. https://doi.org/10.1016/j.fuel.2017.03.001.

[15] M. Gülüm, A. Bilgin, Density, flash point and heating value variations of corn oil biodiesel-diesel fuel blends, Fuel Process. Technol. 134 (2015) 456–464. https://doi.org/10.1016/j.fuproc.2015.02.026.

[16] M. Gülüm, A. Bilgin, Two-term power models for estimating kinematic viscosities of different biodiesel-diesel fuel blends, Fuel Process. Technol. 149 (2016) 121–130. https://doi.org/10.1016/j.fuproc.2016.04.013.

[17] R. Velvarská, A. Vráblík, M. Fiedlerová, R. Černý, Near-infrared spectroscopy for determining the oxidation stability of diesel, biodiesel and their mixtures, Chem. Pap. 73 (2019) 2987–2993. https://doi.org/10.1007/s11696-019-00852-4.

[18] S. Wang, S. Liu, Y. Yuan, J. Zhang, J. Wang, D. Kong, Simultaneous detection of

different properties of diesel fuel by near infrared spectroscopy and chemometrics, Infrared Phys. Technol. 104 (2020) 103111. https://doi.org/10.1016/j.infrared.2019.103111.

[19] S. Liu, S. Wang, C. Hu, S. Zhan, D. Kong, J. Wang, Rapid and accurate determination of diesel multiple properties through NIR data analysis assisted by machine learning, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 277 (2022) 121261. https://doi.org/10.1016/j.saa.2022.121261.

[20] S. Wang, S. Liu, Y. Yuan, J. Zhang, Z. Wang, X. Che, A novel CC-tSNE-SVR model for rapid determination of diesel fuel quality by near infrared spectroscopy, Infrared Phys. Technol. 106 (2020) 103276. https://doi.org/10.1016/j.infrared.2020.103276.

[21] A. Vráblík, R. Velvarská, K. Štěpánek, M. Pšenička, J.M. Hidalgo, R. Černý, Rapid Models for Predicting the Low-Temperature Behavior of Diesel, Chem. Eng. Technol. 42 (2019) 735–743. https://doi.org/10.1002/ceat.201800549.

[22] I. Hradecká, R. Velvarská, K.D. Jaklová, A. Vráblík, Rapid determination of diesel fuel properties by near-infrared spectroscopy, Infrared Phys. Technol. 119 (2021) 103933. https://doi.org/10.1016/j.infrared.2021.103933.

[23] R. Velvarská, A. Vráblík, J.M. Hidalgo-Herrador, R. Černý, Near-infrared spectroscopy to determine cold-flow improver concentrations in diesel fuel, Infrared Phys. Technol. 110 (2020) 103445. https://doi.org/10.1016/j.infrared.2020.103445.

[24] A.B.F. Câmara, L.S. de Carvalho, C.L.M. de Morais, L.A.S. de Lima, H.O.M. de Araújo, F.M. de Oliveira, K.M.G. de Lima, MCR-ALS and PLS coupled to NIR/MIR spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends, Fuel. 210 (2017) 497–506. https://doi.org/10.1016/j.fuel.2017.08.072.

[25] I. Barra, M. Kharbach, E.M. Qannari, M. Hanafi, Y. Cherrah, A. Bouklouze, Predicting cetane number in diesel fuels using FTIR spectroscopy and PLS regression, Vib. Spectrosc. 111 (2020) 103157. https://doi.org/10.1016/j.vibspec.2020.103157.

[26] Z.S. Baird, V. Oja, Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density, Chemom. Intell. Lab. Syst. 158 (2016) 41–47. https://doi.org/10.1016/j.chemolab.2016.08.004.

[27] A.D.V. Máquina, B.V. Sitoe, J.E. Buiatte, D.Q. Santos, W.B. Neto, Quantification and classification of cotton biodiesel content in diesel blends, using mid-infrared spectroscopy and chemometric methods, Fuel. 237 (2019) 373–379. https://doi.org/10.1016/j.fuel.2018.10.011.

[28] L. Alyson, NIR spectroscopy in the petrochemical and refinery industry: The ASTM compliant tool for QC and product screening – Part 1, (2021). https://www.metrohm.com/es_es/discover/blog/20-21/nir-spectroscopy-in-the-petrochemical-and-refinery-industry--the.html.

[29] G.E. Fodor, R.A. Mason, S.A. Hutzler, Estimation of middle distillate fuel properties by FT-IR, Appl. Spectrosc. 53 (1999) 1292–1298. https://doi.org/10.1366/0003702991945542.

[30] American Society for Testing Materials, ASTM E1655-17 Standard Practices for Infrared Multivariate Quantitative Analysis, (2017) 1–29. https://doi.org/10.1520/E1655-17.

[31] O.O. Soyemi, M.A. Busch, K.W. Busch, Multivariate Analysis of Near-Infrared Spectra Using the G-Programming Language, J. Chem. Inf. Comput. Sci. 40 (2000) 1093–1100. https://doi.org/10.1021/ci000447r.

[32] V.O. Santos Jr., 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. 547 (2005) 188–196. https://doi.org/10.1016/j.aca.2005.05.042.

[33] N. Pasadakis, S. Sourligas, C. Foteinopoulos, Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks, Fuel. 85 (2006) 1131–1137. https://doi.org/10.1016/j.fuel.2005.09.016.

[34] J.S. Oliveira, R. Montalvão, L. Daher, P.A.Z. Suarez, J.C. Rubim, Determination of methyl ester contents in biodiesel blends by FTIR-ATR and FTNIR spectroscopies, Talanta. 69 (2006) 1278–1284. https://doi.org/10.1016/j.talanta.2006.01.002.

[35] L. de F. Bezerra de Lira, F.V. Cruz de Vasconcelos, C. Fernandes Pereira, A.P. Silveira Paim, L. Stragevitch, M.F. Pimentel, Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration, Fuel. 89 (2010) 405–409. https://doi.org/10.1016/j.fuel.2009.05.028.

[36] F.B. Gonzaga, C. Pasquini, A low cost short wave near infrared

spectrophotometer: Application for determination of quality parameters of diesel fuel, Anal. Chim. Acta. 670 (2010) 92–97. https://doi.org/10.1016/j.aca.2010.04.060.

[37] M.G. Nespeca, R.R. Hatanaka, D.L. Flumignan, E. De Oliveira, Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis, J. Anal. Methods Chem. (2018) 1795624. https://doi.org/10.1155/2018/1795624.

[38] B. Zhan, J. Yang, Measurement of Diesel Cetane Number Using Near Infrared Spectra and Multivariate Calibration, 100 (2017) 240–247. https://doi.org/10.2991/icmeim-17.2017.41.

[39] 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. 165 (2016) 379–388. https://doi.org/10.1016/j.fuel.2015.10.079.

[40] C.L. Cunha, A.S. Luna, R.C.G. Oliveira, G.M. Xavier, M.L.L. Paredes, A.R. Torres, Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate calibration, Fuel. 204 (2017) 185–194. https://doi.org/10.1016/j.fuel.2017.05.057.

[41] P. de Peinder, T. Visser, D.D. Petrauskas, F. Salvatori, F. Soulimani, B.M. Weckhuysen, Partial least squares modeling of combined infrared, 1H NMR and 13C NMR spectra to predict long residue properties of crude oils, Vib. Spectrosc. 51 (2009) 205–212. https://doi.org/10.1016/j.vibspec.2009.04.009.

[42] M.K. Moro, F.D. dos Santos, G.S. Folli, W. Romão, P.R. Filgueiras, A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy, Fuel. 303 (2021) 121283. https://doi.org/10.1016/j.fuel.2021.121283.

[43] T.I. Dearing, W.J. Thompson, C.E. Rechsteiner, B.J. Marquardt, Characterization of crude oil products using data fusion of process Raman, infrared, and nuclear magnetic resonance (NMR) spectra, Appl. Spectrosc. 65 (2011) 181–186. https://doi.org/10.1366/10-05974.

[44] P. Mishra, F. Marini, A. Biancolillo, J.M. Roger, Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques, Talanta. 223 (2021) 121693. https://doi.org/10.1016/j.talanta.2020.121693.

[45] L.M. de Aguiar, D. Galvan, E. Bona, L.A. Colnago, M.H.M. Killner, Data fusion of middle-resolution NMR spectroscopy and low-field relaxometry using the Common Dimensions Analysis (ComDim) to monitor diesel fuel adulteration, Talanta. 236 (2022) 122838. https://doi.org/10.1016/j.talanta.2021.122838.

[46] J. Buendia-Garcia, M. Lacoue-Negre, J. Gornay, S. Mas-Garcia, R. Bendoula, J.M. Roger, Variable selection and data fusion for diesel cetane number prediction, Fuel. 332 (2023) 126297. https://doi.org/10.1016/j.fuel.2022.126297.

[47] H.A.G. Al-kaf, K.S. Chia, N.A.M. Alduais, A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum, Pet. Sci. Technol. 36 (2018) 411–418. https://doi.org/10.1080/10916466.2018.1425717.

[48] T. Bolanca, S. Marinovic, S. Ukic, A. Jukic, V. Rukavina, A. Chim, Development of Artificial Neural Network Model for Diesel Fuel Properties Prediction using Vibrational Spectroscopy, Acta Chim. Slov. 59 (2012) 249–257.

[49] S. Marinović, T. Bolanča, Š. Ukić, V. Rukavina, A. Jukić, Prediction of diesel fuel cold properties using artificial neural networks, Chem. Technol. Fuels Oils. 48 (2012) 67–74. https://doi.org/10.1007/s10553-012-0339-y.

[50] D.M. Korres, G. Anastopoulos, E. Lois, A. Alexandridis, H. Sarimveis, G. Bafas, A neural network approach to the prediction of diesel fuel lubricity, Fuel. 81 (2002) 1243–1250. https://doi.org/10.1016/S0016-2361(02)00020-0.

[51] B. Basu, G.S. Kapur, A.S. Sarpal, R. Meusinger, A Neural Network Approach to the Prediction of Cetane Number of Diesel Fuels Using Nuclear Magnetic Resonance (NMR) Spectroscopy, Energy and Fuels. 17 (2003) 1570–1575. https://doi.org/10.1021/ef030083f.

[52] C. Wu, J. Zhang, W. Li, Y. Wang, H. Cao, Artificial neural network model to predict cold filter plugging point of blended diesel fuels, Fuel Process. Technol. 87 (2006) 585–590. https://doi.org/10.1016/j.fuproc.2004.07.005.

[53] H. Yang, Z. Ring, Y. Briker, N. McLean, W. Friesen, C. Fairbridge, Neural network prediction of cetane number and density of diesel fuel from its chemical composition determined by LC and GC-MS, Fuel. 81 (2002) 65–74. https://doi.org/10.1016/S0016-2361(01)00121-1.

[54] F.M. De Oliveira, L.S. De Carvalho, L.S.G. Teixeira, C.H. Fontes, K.M.G. Lima, A.B.F. Câmara, H.O.M. Araújo, R. V. Sales, Predicting Cetane Index, Flash Point,

and Content Sulfur of Diesel-Biodiesel Blend Using an Artificial Neural Network Model, Energy and Fuels. 31 (2017) 3913–3920. https://doi.org/10.1021/acs.energyfuels.7b00282.

[55] Y. Zheng, M.S. Shadloo, H. Nasiri, A. Maleki, A. Karimipour, I. Tlili, Prediction of viscosity of biodiesel blends using various artificial model and comparison with empirical correlations, Renew. Energy. 153 (2020) 1296–1306. https://doi.org/10.1016/j.renene.2020.02.087.

[56] M. Hamadache, C. Si-Moussa, M. Laidi, S. Hanini, S. Belmadani, Artificial Neural Network Models for Prediction of Density and Kinematic Viscosity of Different Systems of Biofuels and Their Blends with Diesel Fuel. Comparative Analysis, Kem. u Ind. 69 (2020) 355–364. https://doi.org/10.15255/kui.2019.053.

[57] M. Yari, G.R. Moradi, M. Abdolmaleki, S. Bashiri, Iranian Journal of Chemical Engineering Predicting the Kinematic Viscosity and Cetane Number of Diesel-Biodiesel Blend using Neural Network and Empirical Models, 19 (2023) 81–94. https://doi.org/10.22034/ijche.2023.345114.1441.

[58] Y. Kassem, Adaptive Neuro-Fuzzy Inference System (ANFIS) and Artificial Neural Network (ANN) for Predicting the Kinematic Viscosity and Density of Biodiesel-Petroleum Diesel Blends, Am. J. Comput. Sci. Technol. 1 (2018) 8–18. https://doi.org/10.11648/j.ajcst.20180101.12.

[59] G. Moradi, M. Mohadesi, B. Karami, R. Moradi, Using Artificial Neural Network for Estimation of Density and Viscosities of Biodiesel-Diesel Blends, J. Chem. Pet. Eng. 49 (2015) 153–165. https://doi.org/10.22059/JCHPE.2015.1807.

[60] S. Raghuvaran, B. Ashok, B. Veluchamy, N. Ganesh, Evaluation of performance and exhaust emission of C.I diesel engine fuel with palm oil biodiesel using an artificial neural network, Mater. Today Proc. 37 (2020) 1107–1111. https://doi.org/10.1016/j.matpr.2020.06.344.

[61] S.O. Giwa, S.O. Adekomaya, K.O. Adama, M.O. Mukaila, Prediction of selected biodiesel fuel properties using artificial neural network, Front. Energy. 9 (2015) 433–445. https://doi.org/10.1007/s11708-015-0383-5.

[62] N. BS, A Review on Application of ANN Model for the Prediction of Fuel Properties of Biodiesel, J. Adv. Res. Mech. Eng. Technol. 06 (2019) 32–37. https://doi.org/10.24321/2454.8650.201906.

[63] F. Al-Shanableh, A. Evcil, M.A. Savaş, Prediction of Cold Flow Properties of

Biodiesel Fuel Using Artificial Neural Network, Procedia Comput. Sci. 102 (2016) 273–280. https://doi.org/10.1016/j.procs.2016.09.401.

[64] R. Piloto-Rodríguez, Y. Sánchez-Borroto, M. Lapuerta, L. Goyos-Pérez, S. Verhelst, Prediction of the cetane number of biodiesel using artificial neural networks and multiple linear regression, Energy Convers. Manag. 65 (2013) 255–261. https://doi.org/10.1016/j.enconman.2012.07.023.

[65] E.G. Giakoumis, C.K. Sarakatsanis, Estimation of biodiesel cetane number, density, kinematic viscosity and heating values from its fatty acid weight composition, Fuel. 222 (2018) 574–585. https://doi.org/10.1016/j.fuel.2018.02.187.

[66] A.S. Ramadhas, S. Jayaraj, C. Muraleedharan, K. Padmakumari, Artificial neural networks used for the prediction of the cetane number of biodiesel, Renew. Energy. 31 (2006) 2524–2533. https://doi.org/10.1016/j.renene.2006.01.009.

[67] A.O. Barradas Filho, A.K.D. Barros, S. Labidi, I.M.A. Viegas, D.B. Marques, A.R.S. Romariz, R.M. De Sousa, A.L.B. Marques, E.P. Marques, Application of artificial neural networks to predict viscosity, iodine value and induction period of biodiesel focused on the study of oxidative stability, Fuel. 145 (2015) 127–135. https://doi.org/10.1016/j.fuel.2014.12.016.

[68] C.I. Rocabruno-Valdés, L.F. Ramírez-Verduzco, J.A. Hernández, Artificial neural network models to predict density, dynamic viscosity, and cetane number of biodiesel, Fuel. 147 (2015) 9–17. https://doi.org/10.1016/j.fuel.2015.01.024.

[69] D.B. Marques, A.O. Barradas Filho, A.R.S. Romariz, I.M.A. Viegas, D.A. Luz, A.K.D. Barros Filho, S. Labidi, A.S. Ferraudo, Recent Developments on Statistical and Neural Network Tools Focusing on Biodiesel Quality, Int. J. Comput. Sci. Appl. 3 (2014) 97–110. https://doi.org/10.14355/ijcsa.2014.0303.01.

[70] M. Fernanda Pimentel, G.M.G.S. Ribeiro, R.S. Da Cruz, L. Stragevitch, J.G.A. Pacheco Filho, L.S.G. Teixeira, Determination of biodiesel content when blended with mineral diesel fuel using infrared spectroscopy and multivariate calibration, Microchem. J. 82 (2006) 201–206. https://doi.org/10.1016/j.microc.2006.01.019.

[71] A. Palou, A. Miró, M. Blanco, R. Larraz, J.F. Gómez, T. Martínez, J.M. González, M. Alcalà, Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 180 (2017)

119–126. https://doi.org/10.1016/j.saa.2017.03.008.

[72] H.Z. Msimanga, C.R. Dockery, D.D. Vandenbos, Classification of local diesel fuels and simultaneous prediction of their physicochemical parameters using FTIR-ATR data and chemometrics, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 279 (2022) 121451. https://doi.org/10.1016/j.saa.2022.121451.

[73] S. Marinović, M. Krištović, B. Špehar, V. Rukavina, A. Jukić, Prediction of diesel fuel properties by vibrational spectroscopy using multivariate analysis, J. Anal. Chem. 67 (2012) 939–949. https://doi.org/10.1134/S1061934812120039.

[74] M.F. Ferrão, M.D.S. Viera, R.E.P. Pazos, D. Fachini, A.E. Gerbase, L. Marder, Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions, Fuel. 90 (2011) 701–706. https://doi.org/10.1016/j.fuel.2010.09.016.

[75] D. Huang, H. Zhou, L. Lin, Biodiesel: An alternative to conventional fuel, Energy Procedia. 16 (2012) 1874–1885. https://doi.org/10.1016/j.egypro.2012.01.287.

[76] G. Knothe, K.R. Steidley, Kinematic viscosity of biodiesel fuel components and related compounds. Influence of compound structure and comparison to petrodiesel fuel components, Fuel. 84 (2005) 1059–1065. https://doi.org/10.1016/j.fuel.2005.01.016.

[77] W.B. Zhang, Review on analysis of biodiesel with infrared spectroscopy, Renew. Sustain. Energy Rev. 16 (2012) 6048–6058. https://doi.org/10.1016/j.rser.2012.07.003.

[78] P. Baptista, P. Felizardo, J.C. Menezes, M.J.N. Correia, Multivariate near infrared spectroscopy models for predicting the methyl esters content in biodiesel, Anal. Chim. Acta. 607 (2008) 153–159. https://doi.org/10.1016/j.aca.2007.11.044.

[79] P. Baptista, P. Felizardo, J.C. Menezes, M.J. Neiva Correia, Multivariate near infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40 °C and density at 15 °C of biodiesel, Talanta. 77 (2008) 144–151. https://doi.org/10.1016/j.talanta.2008.06.001.

[80] I.K. de Oliveira, W.F. de Carvalho Rocha, R.J. Poppi, Application of near infrared spectroscopy and multivariate control charts for monitoring biodiesel blends, Anal. Chim. Acta. 642 (2009) 217–221. https://doi.org/10.1016/j.aca.2008.11.003.

[81] R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, S.L. Rose-Pehrsson, Rapid fuel quality surveillance through chemometric

modeling of near-infrared spectra, Energy and Fuels. 23 (2009) 1610–1618. https://doi.org/10.1021/ef800869t.

[82] C. Brouillette, W. Smith, C. Shende, Z. Gladding, S. Farquharson, R.E. Morris, J.A. Cramer, J. Schmitigal, Analysis of Twenty-Two Performance Properties of Diesel, Gasoline, and Jet Fuels Using a Field-Portable Near-Infrared (NIR) Analyzer, Appl. Spectrosc. 70 (2016) 746–755. https://doi.org/10.1177/0003702816638279.

[83] M.C. Breitkreitz, I.M. Raimundo, J.J.R. Rohwedder, C. Pasquini, H.A. Dantas Filho, G.E. José, M.C.U. Araújo, Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration, Analyst. 128 (2003) 1204–1207. https://doi.org/10.1039/b305265f.

[84] I. Pinheiro Soares, T. F. Rezende, I.C. P. Fortes, Determination of sulfur in diesel using ATR/ FTIR and multivariate calibration, Eclet. Quim. 35 (2010) 71–78. https://doi.org/10.26850/1678-4618EQJ.V35.2.2010.P71-78.

[85] 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. https://doi.org/10.1016/j.talanta.2017.07.094.

[86] Q. Zheng, H. Huang, S.P. Zhu, B.H. Qi, X. Tang, Quantitative and qualitative prediction of sulfur content in diesel by near infrared spectroscopy, J. Near Infrared Spectrosc. 31 (2023) 63–69. https://doi.org/10.1177/09670335231153960.

[87] J.C.L. Alves, R.J. Poppi, Simultaneous determination of hydrocarbon renewable diesel, biodiesel and petroleum diesel contents in diesel fuel blends using near infrared (NIR) spectroscopy and chemometrics, Analyst. 138 (2013) 6477–6487. https://doi.org/10.1039/c3an00883e.

[88] T. Fearn, Standardisation and Calibration Transfer for near Infrared Instruments: A Review, J. Near Infrared Spectrosc. 9 (2001) 229–244. https://doi.org/10.1255/jnirs.309.

[89] J.J. Workman, A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy, Appl. Spectrosc. 72 (2018) 340–365. https://doi.org/10.1177/0003702817736064.

[90] P. Mishra, R. Nikzad-langerodi, F. Marini, J. Michel, A. Biancolillo, D.N. Rutledge,

S. Lohumi, Trends in Analytical Chemistry Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers ? The answer is not always, Trends Anal. Chem. 143 (2021) 116331. https://doi.org/10.1016/j.trac.2021.116331.

[91] C.F. Pereira, M.F. Pimentel, R.K.H. Galvão, F.A. Honorato, L. Stragevitch, M.N. Martins, A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers, Anal. Chim. Acta. 611 (2008) 41–47. https://doi.org/10.1016/j.aca.2008.01.071.

[92] B. Perston, N. Harris, Diamond ATR and Calibration Transfer for Biodiesel-Blend Analysis by ASTM D7371, Seer Green, UK, 2009. https://resources.perkinelmer.com/lab-solutions/resources/docs/APP_Biodiesel-BlendAnalysisbyASTMD7371.pdf.

[93] J.B. Cooper, C.M. Larkin, M.F. Abdelkader, Calibration transfer of near-IR partial least squares property models of fuels using virtual standards, J. Chemom. 25 (2011) 496–505. https://doi.org/10.1002/cem.1395.

[94] 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. https://doi.org/10.1016/j.aca.2016.12.018.

[95] R.M. Balabin, S. V. Smirnov, Interpolation and extrapolation problems of multivariate regression in analytical chemistry: Benchmarking the robustness on near-infrared (NIR) spectroscopy data, Analyst. 137 (2012) 1604–1610. https://doi.org/10.1039/c2an15972d.

[96] K.J. Johnson, J. Schmitigal, G.J. Walker, NIR Spectrometry and Fuel Quality Assessment, Washington, United States, 2019. https://apps.dtic.mil/sti/pdfs/AD1089383.pdf.

[97] R.M. Balabin, E.I. Lomakina, R.Z. Safieva, Neural network (ANN) approach to biodiesel analysis : Analysis of biodiesel density , kinematic viscosity , methanol and water contents using near infrared (NIR) spectroscopy, Fuel. 90 (2010) 2007–2015. https://doi.org/10.1016/j.fuel.2010.11.038.

[98] H. Yang, P.R. Griffiths, J.D. Tate, Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase Fourier transform infrared spectra, Anal. Chim. Acta. 489

(2003) 125–136. https://doi.org/10.1016/S0003-2670(03)00726-8.

[99] S. Haykin, Neural Networks - A Comprehensive Foundation, Pearson Education, Inc., Delhi, 2005.

[100] V. Kůrková, Kolmogorov's theorem and multilayer neural networks, Neural Networks. 5 (1992) 501–506. https://doi.org/10.1016/0893-6080(92)90012-8.

[101] J.F. MacGregor, T. Kourti, Statistical process control of multivariate processes, Control Eng. Pract. 3 (1995) 403–414. https://doi.org/10.1016/0967-0661(95)00014-L.

[102] R.P. Cogdill, C.A. Anderson, J.K. Drennen, Process analytical technology case study, part III: Calibration monitoring and transfer, AAPS PharmSciTech. 6 (2005) 284–97. https://doi.org/10.1208/pt060239.

Chapter II
Theoretical foundation

Chapter II

II.1 Infrared spectroscopy

Infrared is the region of the electromagnetic spectrum between 750 and 10^6 nm (13.333-10 cm^{-1}). This region is divided into near-infrared (NIR), mid-infrared (MIR), and far-infrared (FIR) according to radiation energy, the nature of the interaction of radiation with matter, and the instrumental requirements (Table II-1).

Table II-1. Infrared region division. Adapted from [1].

Region	**Wavenumber interval** (cm^{-1})	**Wavelength range** (nm)	**Origin of absorption (transition type)**
NIR	13.333 - 4000	750 - 2500	Overtones and combination bands of fundamental molecular vibrations
MIR	4000 - 400	2500 - 50000	Fundamental molecular vibrations and rotations
FIR	400 - 10	50000 - 10^6	Molecular rotations and skeletal vibration

The spectroscopy associated with IR radiation, which is related to molecular vibrations and rotations, is a powerful analytical tool for the quick and easy characterization of samples, allowing the simultaneous determination of physical and chemical parameters. It is also a valuable tool for the control of industrial processes, especially with instrumental configurations with optical fibers. Among the vibrational spectroscopy techniques, MIR and NIR spectroscopies stand out as widely established techniques used in qualitative and quantitative applications.

NIR spectroscopy is the vibrational spectroscopy in the 13.333 – 4000 cm^{-1} region [2]. The absorption bands in this region are characteristic of the first and second overtones and combination bands of the fundamental vibrations of the functional groups C-H, N-H, and O-H in the mid-infrared [3] (Figure II-1). They are characterized by being wide bands of low intensity [2]. This feature allows the quantitative determination of analytes with high concentrations without pre-treatment of the sample.

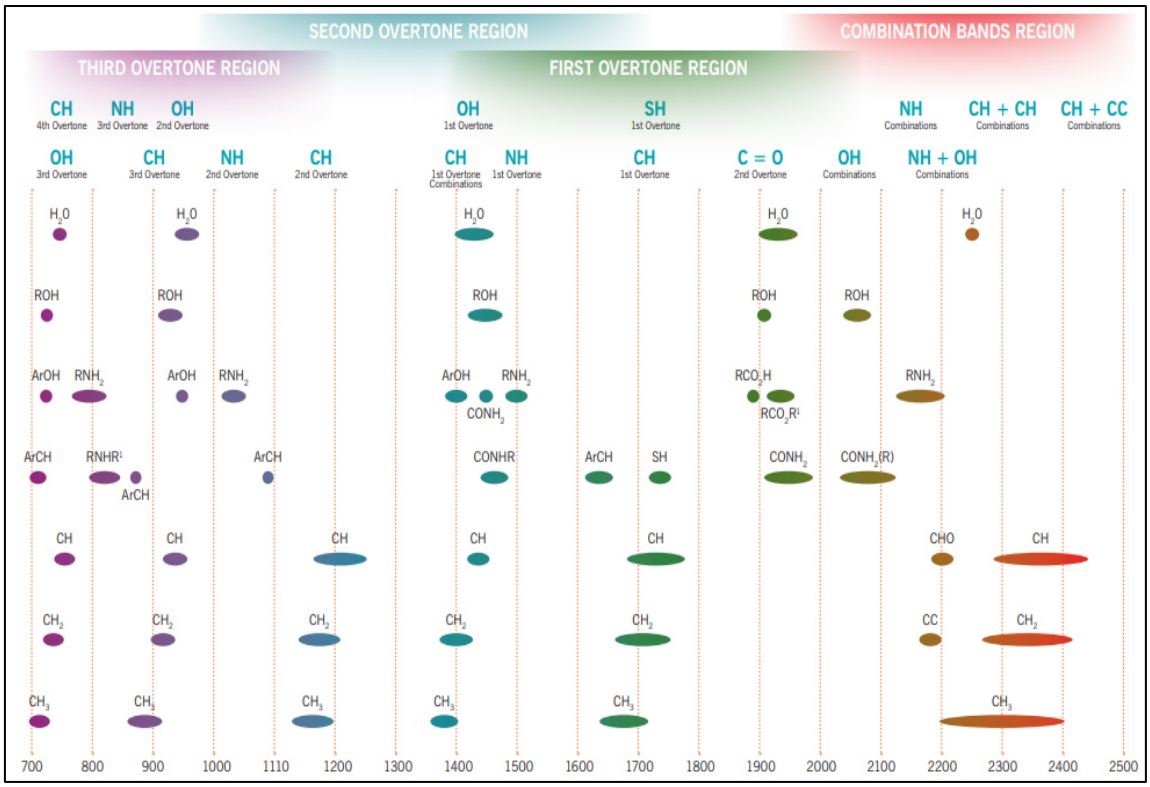


Figure II-1. Overtone and combination NIR band assignment (source: ASD Inc. 2005-2013).

MIR spectroscopy is the vibrational spectroscopy technique in the region 4000 – 400 cm^{-1}. The absorption bands in this region result from transitions between the fundamental vibrational level ($\nu = 0$) and the first excited vibrational level ($\nu = 1$). Simultaneously with vibrational transitions, rotational transitions also occur in this region, which usually overlap with the vibrational bands [1]. Although the MIR region is very useful for structural characterization (Figure II-2), it is less used for online quantitative analysis because the samples must be usually pre-treated due to the high intensity of the absorption bands.

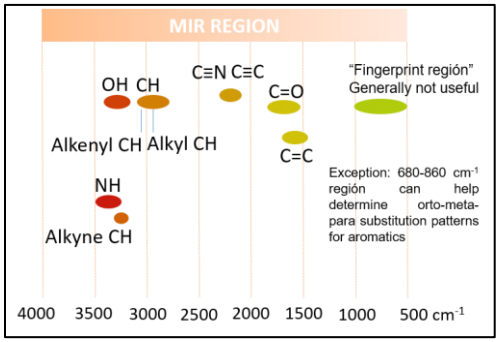


Figure II-2. MIR bands assignment.

There are different modes of measurement in the infrared region [1]. The classical mode for the analysis of liquid samples is transmittance (Figure II-3).

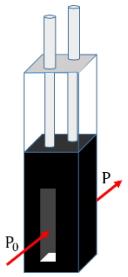


Figure II-3. Near-infrared transmittance.

The transmittance at a given wavenumber $\tilde{v}$ is given by:

$$T = \frac{P}{P_0} \tag{II.1}$$

where P is the radiant power emerging from the sample when a beam of radiant power P_0 reaches the sample through the transparent cuvette of length l that contains the sample. Absorbance is defined as $-\log(\mathrm{T})$ and relates to concentration through Beer-Lambert's law (eq. II.2) [1,2]

$$\mathrm{A} = -\log(\mathrm{T}) = \log\left(\frac{\mathrm{P_0}}{\mathrm{P}}\right) = \varepsilon l c \tag{II.2}$$

where ε is the molar absorptivity, l is the optical path, and c is the concentration of the absorbent species.

II.2 Exploratory analysis

Due to the overlap of spectral bands and complexity of IR spectra, multivariate analysis is a crucial player in the applications of IR spectroscopy, used in exploratory studies, classification, and quantification [4–6]. Exploratory data analysis is performed with unsupervised methods, where principal component analysis (PCA) is possibly the most widely used [7–9].

PCA is a factorization method that decomposes a matrix $\mathbf{X}$ $(I \times K)$ into the product of two matrices according to:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{II.3}$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$ $(I \times A)$ is the scores matrix and has orthogonal columns, $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_A]$ $(K \times A)$ is the loadings matrix and has orthogonal and normalized columns, $\mathbf{E}$ $(I \times K)$ is the matrix of residuals, and A is the number of retained factors.

PCA projects the multivariate data into a reduced space of principal components (PCs), retaining as much information as possible from the original space [10] (Figure II-4). PCs are linear combinations of the original variables and are orthogonal to each other. They are ordered according to the variance they explain, with PC1 accumulating the maximum spectral variability of the I samples at K wavelengths. PCA, among other applications, allows the visualization of similarities between samples and the identification of spectral outliers [11–14].

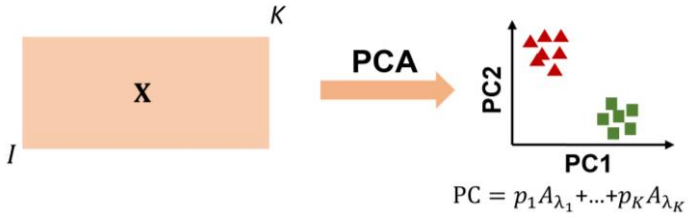


Figure II-4. Principal component analysis representation.

T-Distributed Stochastic Neighbor Embedding (t-SNE) is another exploratory technique that is used to visualize local clusters and patterns in high-dimensional data [15]. Conversely to PCA, t-SNE is a nonlinear dimensionality reduction technique designed to preserve only local similarities between data points. This algorithm minimizes the Kullback-Leibler (KL) divergence between a Gaussian probability distribution used to model the similarity between points in the original space and other similar probability distributions over the points in the lower-dimensional space, ensuring that similar points in the high-dimensional space remain close in the lower-dimensional space.

II.3 Linear multivariate calibration model

The quantitative analysis of a sample using its infrared spectrum involves a multivariate calibration model that relates the property value and the spectrum. Inverse multivariate calibration is based on developing a quantitative model for predicting a property of interest based on predictor variables. In this thesis, multivariate calibration models are used to predict a physicochemical property of a diesel sample ($\hat{y}_{\text{unk}}$) from the vector of measured spectra $\mathbf{x}_{\text{unk}}^{\text{T}} = [x_{\text{unk},1}, \ldots, x_{\text{unk},K}]$ as follows:

$$\hat{y}_{\text{unk}} = b_0 + x_{\text{unk},1} b_1 + \cdots + x_{\text{unk},K} b_K = b_0 + \sum_{k=1}^{K} x_{\text{unk},k}\, b_k \qquad \text{(II.4)}$$

where b_0 is the constant term, $x_{\text{unk},k}$ is the absorbance at sensor *k* and b_k is the regression coefficient for sensor *k*.

II.3.1 PLS algorithm

PLS regression is perhaps the most commonly used model for calibration with IR spectra [16]. PLS regression can handle the multicollinearity in $\mathbf{X}$ by creating latent variables (LVs) that explain most of the variability in the response variables $\mathbf{Y}$ [17,18]. The PLS1 regression based on the non-iterative partial least squares (NIPALS) algorithm consists of two outer relations (X and y-block) and an inner relation between both blocks [19] (Figure II-5). The outer relation for each block is based on the decomposition of $\mathbf{X}$ and $\mathbf{y}$ to obtain a space of LVs:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad \text{(II.5)}$$

$$\mathbf{y} = \mathbf{u}q + \mathbf{f} \qquad \text{(II.6)}$$

where $\mathbf{X}$ is the matrix of predictor variables, $\mathbf{y}$ is the response vector, $\mathbf{T}$ and $\mathbf{u}$ are scores, $\mathbf{P}$ and q are the loadings, and $\mathbf{E}$ and $\mathbf{f}$ are the residuals of $\mathbf{X}$ and $\mathbf{y}$. The inner relation is made by regressing $\mathbf{u}$ against $\mathbf{t}$ for every LV *a*.

$$\hat{\mathbf{u}}_a = b_{\text{inner},a}\mathbf{t}_a \qquad \text{(II.7)}$$

Finally, the predicted value $\hat{y}$ is estimated as follows:

$$\hat{y} = \mathbf{x}^{\mathrm{T}}\mathbf{b} \qquad \text{(II.8)}$$

where the vector of regression coefficients is $\mathbf{b} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{b}_{\text{inner}}$ and $\mathbf{W}$ is the matrix of weights for the X-block.

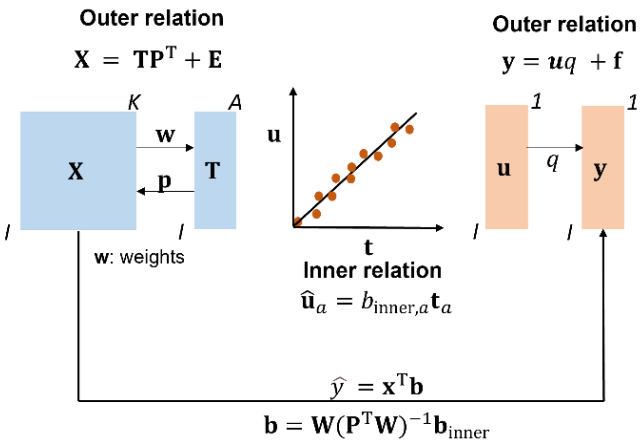


Figure II-5. Stages of PLS1 regression. Adapted from [20,21].

II.3.2 Establishment of a PLS calibration model

The establishment of a calibration model involves the following stages: design and preparation of samples, recording of spectra and determination of the property of interest using the reference method, pretreatment of data, calculation and validation of the model, application of the model, and model maintenance (Figure II-6).

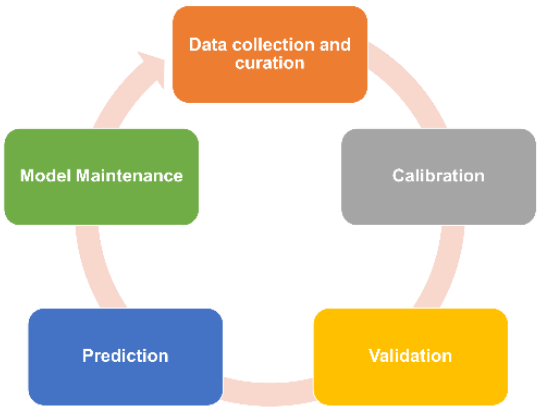


Figure II-6. Stages in the use of a multivariate calibration model.

II.3.2.1 Data collection and curation

Data collection and curation involves collecting and preparing the samples, determining the reference values of the properties, the acquisition of the spectra, and the data splitting.

Available data are split into training, validation, and test sets. The training set is used to calculate the model coefficients. It must be large enough to cover all the variability of the samples to be predicted so that the model can predict accurately without the need for extrapolation. The validation set is used to optimize the customizable parameters of the model (e.g., number of latent variables). The test set (a.k.a. prediction set) is used to evaluate the actual predictive ability of the model with samples that have been neither used to calculate the model coefficients nor used to choose one model among other competing models [22]. When the dataset is not large enough, the validation set is often omitted, and cross-validation performs this function.

Data splitting methods act as a fundamental step prior to the data processing step [23]. Some algorithms that can be used for sample selection include the Kennard-Stone (KS) algorithm [24], leverage-based algorithms [25], or random selection. Due to its simplicity, random selection is one of the most used.

II.3.2.2 Calibration

- Data Preprocessing

Before calculating the model coefficients, it is usually advantageous to preprocess the spectra to remove signal variability that is not related to the property of interest. A common pre-processing is mean-centering, which involves subtracting the average spectrum of the dataset from each spectrum (eq. II.9):

$$x_{ik}^{mean-centered} = x_{ik} - \bar{x}_k \quad \text{(II.9)}$$

where, x_{ik} is the absorbance of the sample i at wavelength k and $\bar{x}_k$ is the average of the absorbances of the training data at wavelength k. This removes the part of the spectrum that is common to all samples and, therefore not useful for prediction.

The first derivative is another common preprocessing, which is often used to accentuate variations in the data and enhance features of interest. In its simplest form, the first derivative is computed by subtracting from the absorbance of the sample i at k wavelength, x_{ik}, the absorbance of the sample at the $k+\delta$ immediate neighboring wavelength, $x_{ik+\delta}$ (eq. II.10) and dividing by the increment of wavelengths.

$$x_{ik}^{1st\ derivative} = \frac{x_{ik} - x_{ik+\delta}}{\Delta_k} \quad \text{(II.10)}$$

Since eq. II.10 increases the noise in the data the first derivative is coupled with the Savitzky-Golay smoothing to reduce the noise while preserving the features of interest.

- Variable selection

Variable selection seeks to improve the predictive ability, computational efficiency, and interpretability of the model [26]. The goal is to identify the subset of wavelengths that produces the smallest error prediction. Some common methods for spectral selection used with PLS are based on a genetic algorithm, interval selection, or a moving window [27–29]. The moving window approach, which is a classical wavelength interval selection method, uses moving window partial least squares regression (MWPLSR) for spectral mapping, providing stable and fast results [30]. A new model is obtained for each continuous size window displacement and tested using cross-validation. There are other modifications of this approach, such as the changeable size moving window partial least squares (CSMWPLS), the searching combination moving window partial least squares (SCMWPLS), and modified changeable size moving window partial least

squares (MCSMWPLSR), between others, has been widely used [31,32]. Figure II-7 shows some modifications of the moving windows approach.

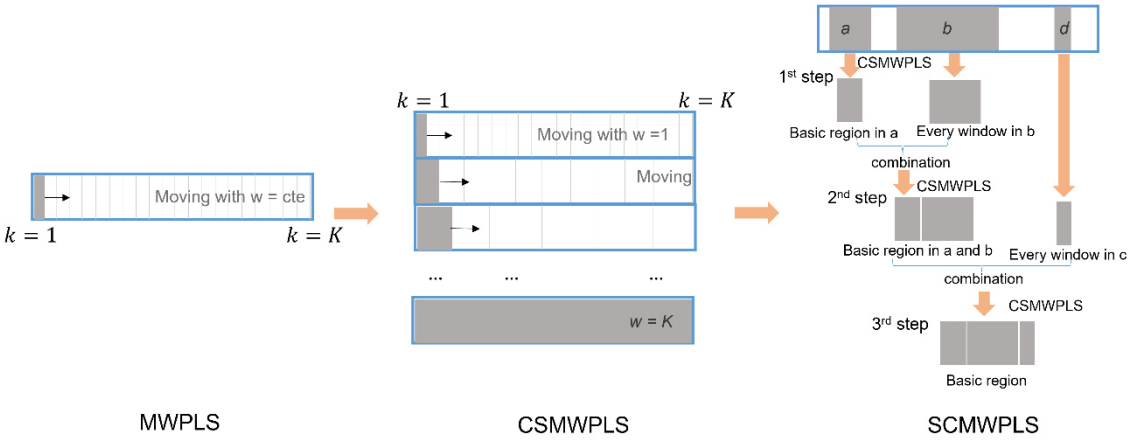


Figure II-7.Some approaches of moving window PLS regression. Adapted from [27].

- Selection of optimal number of latent variables in PLS

A key step in the calculation of a PLS model is the decision of the number of LVs that are needed to retain the spectral variability representative of the samples. Too a high number of LVs creates overfitted models with a loss of generalization ability and less predictive performance. Too a low number of LVs leads to under-fitted models whose predictions are mostly biased. Cross-validation (CV) is one of the most used methods for determining the optimal number of LVs [33]. During cross-validation, a subset of samples is removed from the dataset and used to evaluate a PLS model that has been calculated without that subset. This step is repeated with the next subset of samples until all have been left out once. Ideally, the optimal number of LVs corresponds to the model with the lowest root mean square error of CV (RMSECV). In practice, when there is not a clear global minimum, softer rules such as “the beginning of a plateau” or “the first local minimum” are used (Figure II-8) [16,33,34].

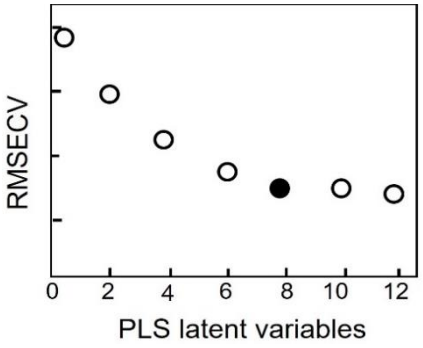


Figure II-8. RMSEV as a function of PLS components. Adapted from [33].

II.3.2.3 Validation

Figure *II-9* shows the flow diagram of the calibration and validation processes [27]. Validation evaluates the predictive ability of the model over a set of validation samples that have not been used to train the model. The root mean square error is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I}(\hat{y}_i - y_i)^2}{I}} \qquad \text{(II.11)}$$

where y_i and $\hat{y}_i$ are the reference value of the property for sample *i* and the value predicted by the model, respectively, and *I* is the number of predicted samples. This value can be either calculated by cross-validation (RMSECV) or using an external validation set (RMSEV) [15,22] and is an estimate of the average prediction error that can be expected when predicting future samples [37].

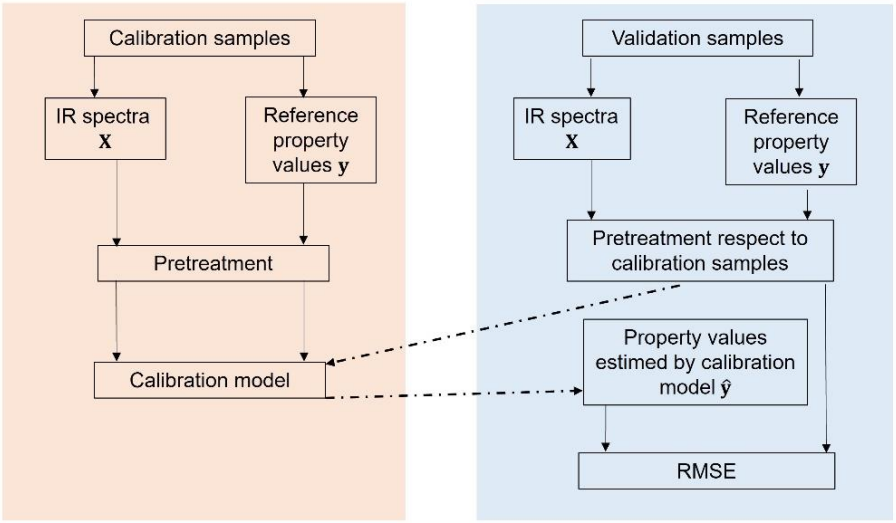


Figure II-9. Flow diagram of calibration and validation. Adapted from [27].

It is common to plot the predicted values $\hat{y}_i$'s against the reference values y_i's where the prediction errors can be observed as vertical deviations from the diagonal of the graph. The coefficient

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} \qquad \text{(II.12)}$$

is then reported as a summary of the performance of the model. To verify if the predictions agree with the values provided by the reference method [38–42], one compares the prediction error with the reproducibility limit of the reference method, R, as described in eq. II.13.

$$-R < \hat{y}_i - y_i < R \tag{II.13}$$

where, y_i is the reference value and $\hat{y}_i$ is the prediction from the PLS model. The reproducibility of a reference method is calculated as the interlaboratory measurement error at a 95% confidence level.

According to the ASTM-E-1655 norm, which outlines practices for infrared multivariate quantitative analysis, the model predictions agree with the reference method if the range specified by eq. II.13 for a specific property encompasses 95% or more of the validation set.

II.3.2.4 Prediction

At the prediction step, the model is used to estimate the property values of incoming samples. Hence, it is essential to adequately establish the limit of applicability of the model to ensure the reliability of the predictions and to address predictive outliers effectively if they are detected. Predictive outliers are samples that fall outside the model's applicability limits. Some reasons can be instrumental failures, new interferences, or errors in spectral acquisition.

- Setting applicability limits of a PLS model

The applicability limits of a calibration model are an important quality parameter of analytical methods that involves the examination of the similarity of an incoming sample to the samples in the calibration set [43]. For this purpose, the leverage, as a measure of the position of a sample in the space of latent variables, and the spectral residuals, as a measure of the orthogonal distance from a sample to the model from selected latent variables, can be used [44]. Specifically, Hotelling's T^2 and the Q statistic as a measure of leverage and spectral residuals, respectively, are commonly used to establish the applicability limits of a PLS model.

- Hotelling's T^2 [45] indicates how far a sample is from the center of the model. For the calibration samples, it is calculated as:

$$\mathrm{T}_i^2 = \sum_{a=1}^{A} \frac{t_{ia}^2}{s_{ia}^2} \quad \text{(II.14)}$$

where t_{ia} is the score of the *i*th sample in the *a*th LV used in the model and s_{ia}^2 is the estimated variance of $\mathbf{t}_a$.

- Q-residuals [46] are a lack-of-fit statistic that measures the spectral residual between a sample spectrum and its projection onto the LVs of the model. It is calculated as follows for sample i:

$$Q_i = \mathbf{e}_i\mathbf{e}_i^{\mathrm{T}} = \mathbf{x}_i\left(\mathbf{I} - \mathbf{P}_{\mathrm{A}}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}}\right)\mathbf{x}_i^{\mathrm{T}} \quad \text{(II.15)}$$

where, $\mathbf{e}_i$ is the spectral residual, $\mathbf{P}_{\mathrm{A}}$ is the matrix of loadings and $\mathbf{x}_i$ is the response vector for sample i.

A decision regarding the similarity of an incoming sample to the training samples can be determined by the values of Q_{unk} and T_{unk}^2 compared to the control limits T_{lim}^2 and Q_{lim}. These control limits can be established by assuming statistical distributions such as F-distribution and a χ^2-distribution for Hotelling's T^2 and *Q*-residuals, respectively. Samples with high Hotelling's T^2 $(T^2 > T_{\mathrm{lim}}^2)$ are at the extremes of the experimental domain and thus, their prediction cannot be trusted. Similarly, a sample with a high Q-value ($Q_{unk} > Q_{lim}$) contains spectral variability not contemplated by the model that may increase the bias in the predictions. Note that high Q and T^2 values for a new sample do not automatically (except for the most extreme cases) imply high prediction errors; it only suggests that the predictions are unreliable [47].

Consequently, to decide the action to be taken for a sample whose spectrum is beyond T_{lim}^2 or Q_{lim}, the property value must be determined using the reference method. Then, if the absolute prediction error is lower than the tolerance limits (R) admitted by the reference method, the sample can be added to the calibration set for the model maintenance step. If the prediction error is higher than R, the prediction sample is a spectral outlier and is rejected. This general procedure for prediction samples is shown in Figure II-10.

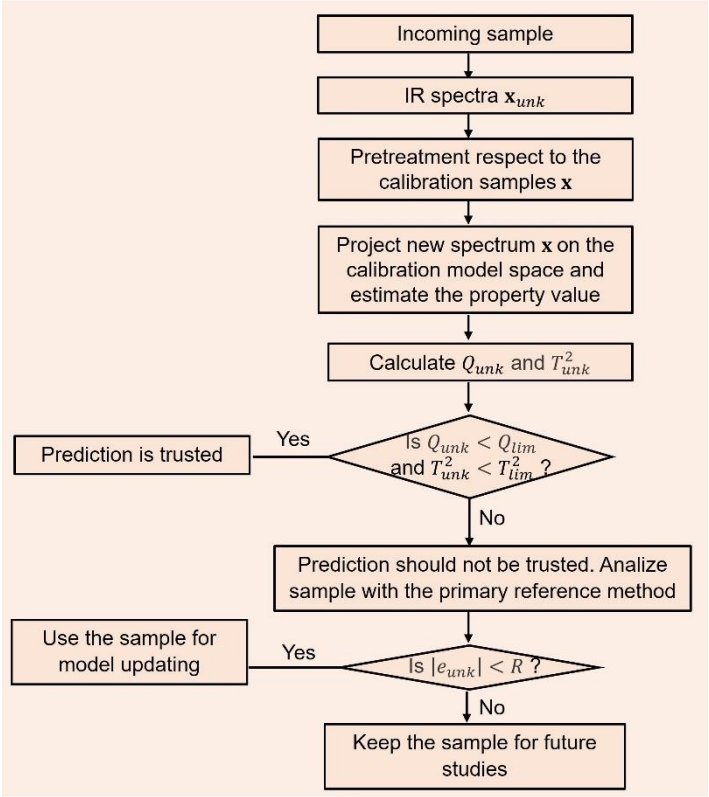


Figure II-10. Procedure of the prediction step. Adapted from [43].

II.3.2.5 Model maintenance

A calibration model should be robust enough to be used for long periods of time. Changes in the instrumental response ("instrumental drift") and changes in the physical-chemical composition of the samples ("product drift") over time [48] can translate into new sources of variability in the spectra not considered during calibration, and that may result in invalid predictions.

Model maintenance refers to the processes for preserving or improving the predictive ability over time and adapting the model to varying working conditions [49] with the least effort and cost. Figure II-11 shows a general scheme of calibration maintenance to identify the new source of variations in the data and to implement strategies for adapting the model, ensuring its continued validity over time.

- Monitoring stability over time

A common practice to detect unexpected or abnormal ranges of variation in infrared measurements over time associated with changes in the environment or instrument (i.e. replacement, repair, or aging of the instrument) is to periodically compare the PLS predictions with the values provided by the reference method. To do this, the spectrum

of a reference sample is periodically measured, and its prediction is monitored. Systematic trends in the predictions over time may indicate unusual variability in the infrared measurements that affect the predictions of all future samples.

Strategies to detect unusual characteristics in the incoming samples (i.e. the raw materials of new batches differ), which do not require the use of the reference method, include the use of multivariate statistical process control (MSPC) rules [46]. These rules allow monitoring of diagnostic measures such as Hotelling's T^2 and Q-residuals, facilitating the identification of their trends toward unacceptable values [47].

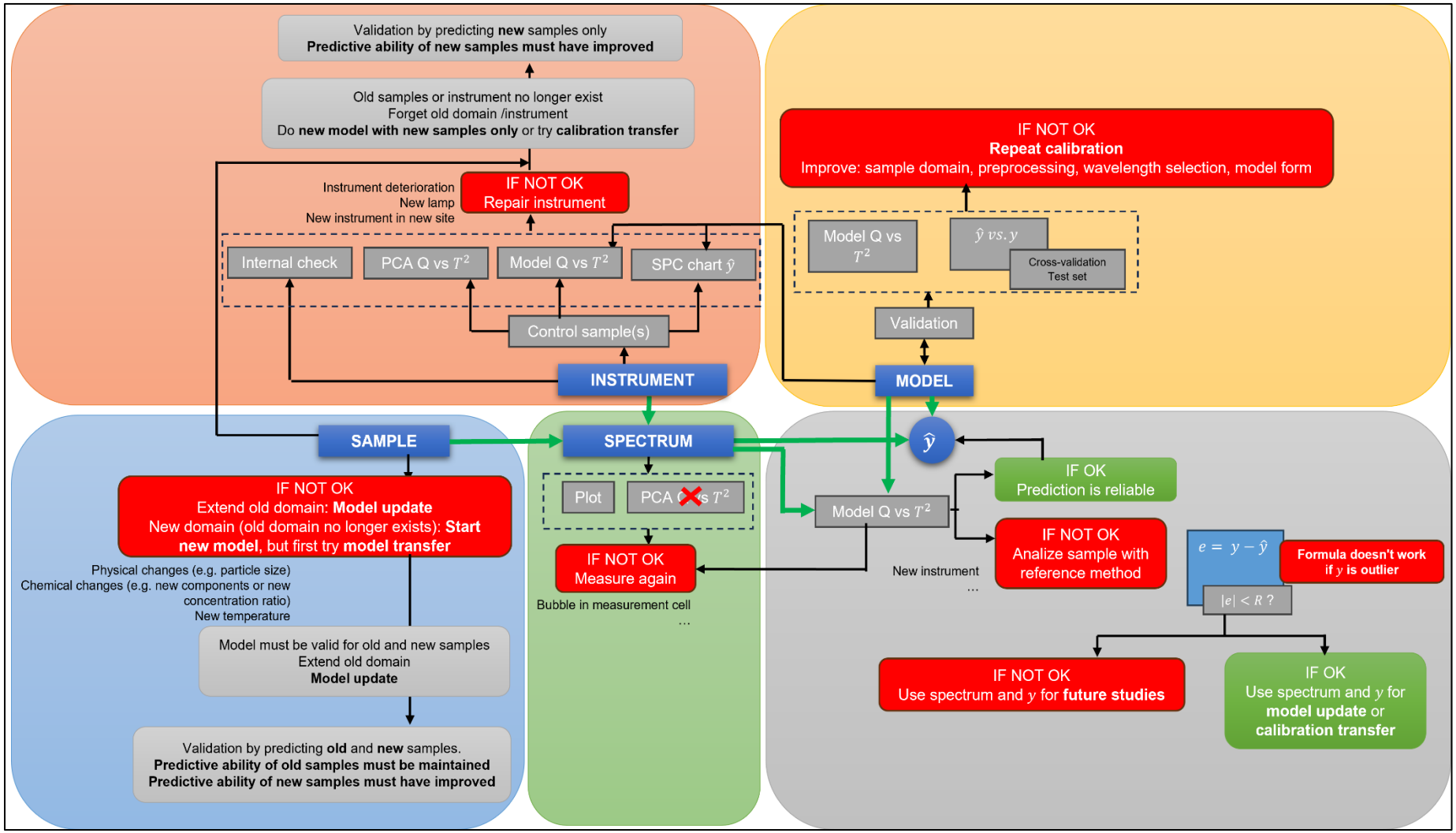


Figure II-11. Calibration maintenance scheme

- Calibration transfer

Calibration transfer reduces discrepancies in instrumental responses or predicted values between two instruments, ensuring that the model's predictions remain accurate and reliable when applied to a new instrument [48,50,51]. In this thesis, we consider transferring a calibration model developed in the source instrument to make it applicable to another (refer to section VI.3).

One way to do this is by measuring a small number of samples, known as transfer standards, on both instruments and then estimating a transfer function between them that can be used to transfer the model to the new instrument. However, this procedure may not always be feasible, e.g., if the source instrument is no longer available or stable calibration standards are unavailable. Alternative standard-free procedures have been developed in recent years [52,53], such as domain adaptation techniques (including transfer component analysis (TCA) [54–58], scatter component analysis (SCA) [59]), domain invariant PLS regression (di-PLS) [60], parameter-free calibration enhancement (PFCE) framework [61]) and the dynamic orthogonal projection (DOP) [56–58,62]. Among these methods, di-PLS and DOP stand out for their versatility. The theoretical background of these two calibration transfer approaches used in this thesis to adapt a PLS model between two instruments is described below.

Domain invariant partial least squares

Domain invariant PLS regression (di-PLS) is a calibration transfer method [60,63–66] that consists of an extended PLS regression with a domain regularization term. This term helps to minimize the variability between the source spectra $\mathbf{X}_\mathrm{S}$ and the target spectra $\mathbf{X}_\mathrm{T}$, while maximizing the covariance between $\mathbf{X}_\mathrm{S}$ and the corresponding response variable $\mathbf{y}_\mathrm{S}$. The di-PLS method is implemented as follows:

In the first step, the NIPALS algorithm is used to extract domain-invariant latent variables between $\mathbf{X}_\mathrm{S}$ $(N_S \times K)$ and $\mathbf{X}_\mathrm{T}$ $(N_T \times K)$. This is achieved by minimizing the function given in eq. II.16. The first term corresponds to the NIPALS objective function, and the second term corresponds to a domain regularization term, which represents an upper limit on the absolute difference between the variances of $\mathbf{X}_\mathrm{S}$ and $\mathbf{X}_\mathrm{T}$ in the direction of $\mathbf{w}$.

$$\min_{\mathbf{w}} \left\| \mathbf{X}_\mathrm{S} - \mathbf{y}_\mathrm{S}\mathbf{w}^\mathrm{T} \right\|_\mathrm{F}^2 + \gamma \mathbf{w}^\mathrm{T} \Lambda \mathbf{w} \qquad \text{(II.16)}$$

where $\| \|_F$ is the Frobenius norm, γ is the domain regularization parameter, $\mathbf{w}$ is a weight vector, $\Lambda = \mathbf{K}\mathrm{diag}(|\lambda_1|, \ldots, |\lambda_k|)\mathbf{K}^T$ is the matrix obtained by calculating the absolute value of the eigenvalues $\lambda_1, \ldots, \lambda_k$ in the eigen decomposition described in eq. II.17 and $\mathbf{K}$ is the eigenvector matrix of the difference between the covariance matrices of $\mathbf{X}_S$ and $\mathbf{X}_T$.

$$\mathbf{K}\mathrm{diag}(|\lambda_1|, \ldots, |\lambda_k|)\mathbf{K}^T = \frac{1}{N_S - 1}\mathbf{X}_S^T\mathbf{X}_S - \frac{1}{N_T - 1}\mathbf{X}_T^T\mathbf{X}_T \quad \text{(II.17)}$$

For each latent variable, the regularization parameter γ in eq. II.16 is determined using the heuristic approach described in reference [64], where the same weight vector was used in both terms of eq. II.16.

The solution of eq. II.16 is obtained by dividing the weight vector $\mathbf{w}$ from eq. II.18 by $\|\mathbf{w}^T\mathbf{w}\|$.

$$\mathbf{w}^T = \frac{\mathbf{y}_S{}^T\mathbf{X}_S}{\mathbf{y}_S{}^T\mathbf{y}_S}(\mathbf{I} + \frac{\gamma}{\mathbf{y}_S{}^T\mathbf{y}_S}\Lambda)^{-1} \quad \text{(II.18)}$$

The scores $\mathbf{t}_S$ and $\mathbf{t}_T$ of the domain-invariant projections corresponding to the direction $\mathbf{w}$ are computed as in eq. II.19

$$\mathbf{t}_S = \mathbf{X}_S\mathbf{w} \text{ and } \mathbf{t}_T = \mathbf{X}_T\mathbf{w} \quad \text{(II.19)}$$

The orthogonalization step to eliminate the variation in the data that is captured by the current LV is described in eq. (II.20).

$$\mathbf{X}_S = \mathbf{X}_S - \mathbf{t}_S(\mathbf{t}_S^T\mathbf{t}_S)^{-1}\mathbf{t}_S^T\mathbf{X}_S \text{ and } \mathbf{X}_T = \mathbf{X}_T - \mathbf{t}_T(\mathbf{t}_T^T\mathbf{t}_T)^{-1}\mathbf{t}_T^T\mathbf{X}_T \quad \text{(II.20)}$$

The remaining steps that follow are the same as the standard algorithm used in PLS regression [63,65].

Dynamic orthogonal projection

Dynamic orthogonal projection (DOP) is a calibration transfer technique used in spectroscopic modeling to address the influences of physical, chemical, and environmental factors [56–58,62]. DOP involves the correction of the calibration dataset by incorporating new reference measurements obtained under the new conditions. This correction is achieved through orthogonal projections defined by the differences between

the source calibration spectra $\mathbf{X}_S$ and the target spectra measured under the new conditions $\mathbf{X}_T$ .The DOP method is implemented as follows:

The virtual standard spectra $\widehat{\mathbf{X}}_T$, the target spectra that should have been measured with the first spectrophotometer are estimated as a linear combination of $\mathbf{X}_S$ following eq. II.21.

$$\widehat{\mathbf{X}}_T = \mathbf{A}\mathbf{X}_S \tag{II.21}$$

The coefficients (a_{im}) of the linear combination $\mathbf{A}$ are calculated by kernel functions centered on $\mathbf{y}_T$, the reference values of $\mathbf{X}_T$, and applied to $\mathbf{y}_S$, the reference values of $\mathbf{X}_S$ as follows: $a_{im} = \mathrm{F}_{y_{Ti}}(\mathbf{y}_{Sm})$, where i and m are the index of the spectra measured in spectrometers 1 and 2, respectively, and $\mathrm{F}_{y_{Ti}}$ is a kernel function. Specifically, due to their continuity properties and the normal distribution assumed for the reference values $\mathbf{y}_S$, the Gaussian kernel was chosen in this thesis for the function $\mathrm{F}_{y_{Ti}}$ following eq. II.22:

$$a_{im} = \frac{1}{\varepsilon\sqrt{2\pi}\sigma(\mathbf{y}_S)} \exp\left(\frac{-(\mathbf{y}_{Ti} - \mathbf{y}_{Sm})^2}{2\varepsilon^2\sigma^2(\mathbf{y}_S)}\right) \tag{II.22}$$

where the parameter ε was tested using the following sequence of values $\varepsilon \in [10^{-4}; 10^{-3.8}; \ldots; 10^{-0.2}; 1]$ and σ is the Gaussian kernel width.

The difference spectra matrix **D** between $\mathbf{X}_T$ and $\widehat{\mathbf{X}}_T$ is calculated as:

$$\mathbf{D} = \mathbf{X}_T - \widehat{\mathbf{X}}_T \tag{II.23}$$

An orthonormal basis **P** of the space spanned by **D** is estimated by principal component analysis:

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{II.24}$$

where $\mathbf{T}$, $\mathbf{P}$, and $\mathbf{E}$ are the scores of $\mathbf{D}$, the corresponding loadings vectors, and the residuals, respectively. The dimension *p* for $\mathbf{P}$ is selected taking into account the percentage of variance explained by PCA.

The calibration spectra are orthogonally projected onto the basis $\mathbf{P}$ to obtain the corrected source data $\mathbf{X}_S^*$:

$$\mathbf{X}_S^* = \mathbf{X}_S(\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathbf{T}}) \tag{II.25}$$

A new PLS regression is performed using $\mathbf{X}_S^*$ and the source reference values $\mathbf{y}_S$. The correction applied to the source calibration database through orthogonal projection is integrated into the model itself. As a result, there is no need to correct new spectra acquired under other scenarios (the target instrument) when using the recalculated model.

- Model updating

Calibration models are updated to expand the calibration space for incoming samples that have a new composition or when the response of the instrument changes [47,67]. The simplest MU strategy involves recalculating the model by adding new spectra to the existing training set. This ensures that the model's predictive ability comprises both old and new conditions.

II.4 Artificial neural networks

Artificial neural networks (ANN) can be used to model non-linear and complex relationships between input features $\mathbf{X}$ and the output $\mathbf{y}$ without any explicit assumptions about the model [68].

II.4.1 Feed-forward neural network

The feed-forward multilayer perceptron or feed-forward neural network (FFNN) is likely the most commonly employed ANN in quantitative analysis using spectroscopic data [35,39,69–72]. In this network, neurons are arranged in layers, and the flow of information occurs in one direction - from the input layer to the output layer [73]. The architecture of a FFNN consists of one input layer, one or more hidden layers, and one output layer (Figure II-12).

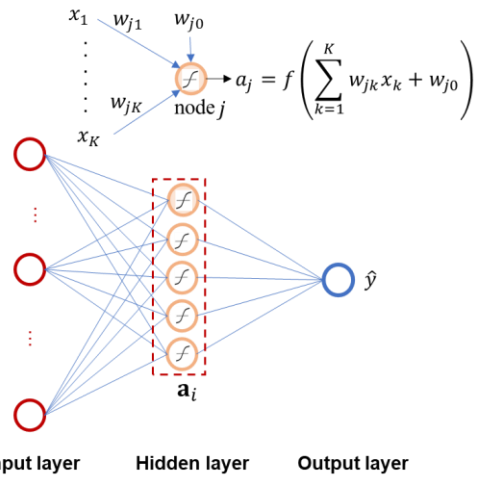


Figure II-12. The architecture of a feed-forward multilayer perceptron with one hidden layer.

The input layer contains all the elements of the spectrum $\mathbf{x} = [x_1, \dots x_k, \dots x_K]^{\mathrm{T}}$. For the first hidden layer, the input of a single node is a linear combination of the input layer (x_k elements of the spectrum $\mathbf{x}$) and f is a nonlinear activation function. This function performs the nonlinear distortion from input to output, thereby constructing a nonlinear model. There are various activation functions such as hyperbolic tangent, sigmoid, and linear function, which produce different bounded outputs. The choice of the activation function depends on the specific problem to be solved and the characteristics of the data [74]. The layer of j nodes processes data from the previous layer according to eq. II.26. Each node in this layer represents a specific kind of feature in the input spectrum.

$$a_j = f\left(\sum_{k=1}^{K} w_{jk} x_k + w_{j0}\right) \tag{II.26}$$

where, a_j is the output of node j of the layer, w_{jk} and w_{j0} are the weights and biases associated with this node, x_k's are the inputs of node j and f is the activation function. In the output layer, which consists of a single node, the x_j's are the outputs from the previous hidden layer, the activation function f is linear and the output is the estimated value for the property of interest, $\hat{y}$.

The weights and biases are estimated by backpropagation from the pairs of spectrum and property values $\{\mathbf{x}, y\}$ of the training set [73,75]. To achieve this, the backpropagation algorithm, which is based on gradient descent, minimizes the loss function given by the mean square error (MSE) [68]:

$$MSE = \frac{1}{I}\sum_{i=1}^{I}(y_i - \hat{y}_i)^2 \tag{II.27}$$

where I is the number of training samples, $\hat{y}_i$ is the predicted property value and $\hat{y}_i$ is the reference property value of the *i*th training sample. The initial values for the weights are random and close to zero [-1,1], so the model starts out almost linear and becomes nonlinear as the weights are optimized [76,77].

II.4.2 Establishment of a FFNN model

The implementation of ANNs can be divided into three steps: data preprocessing, learning, and testing.

II.4.2.1 Data preprocessing

Before training the network, it is also useful to apply normalization or standardization to scale the input values ($\mathbf{x}, y$) so that they are within a certain range [–1,1] or [0,1]. Min-max normalization is one of the most used data normalization methods in ANNs since it provides to the network with a feasible manner to assign the correct weights to network inputs by preventing certain features from dominating the learning process due to differences in their scales [78,79]. This pretreatment scales x_{ik} of $\mathbf{x}$ in the range [MIN, MAX] according to eq. II.28.

$$x_{ik}^{'} = \frac{x_{ik} - \min(\mathbf{x}_k)}{\max(\mathbf{x}_k) - \min(\mathbf{x}_k)}(MAX - MIN) + MIN \quad \text{(II.28)}$$

where, MAX is the new maximum of variable k (in this thesis $MAX = 1$), MIN is the new minimum of variable k (in this thesis $MIN = -1$), x_{ik} is the absorbance value of sample *i* at sensor *k*. The y-variable is pretreated likewise. This step will accelerate subsequent calculations (learning step) and improve the training efficiency.

II.4.2.2 Learning step

During the learning step, the abstraction capacity of the model is developed based on two processes carried out in parallel: training and validation. To achieve this, the model is fine-tuned using both the training and validation sets until the desired level of output precision is attained. The training set is used by the network to learn the spectral pattern, while the validation set is used to evaluate the generalization ability of the trained network [80] to prevent “overfitting”, ensuring that the model performs well not just on the trained data but also on new data [81]. Overfitting occurs when the training error decreases and the validation error starts to increase [82]. Hence, the training process stops when the lowest validation error is obtained.

The learning process constitutes a multivariable optimization problem that may encompass hundreds of variables [84]. Among the most crucial are the number of hidden layers, the number of nodes in the hidden layer, the type of regularization, and the network weights.

Selection of the number of nodes and hidden layers

In the majority of prediction problems, employing a single hidden layer is often sufficient [84]. This is due to the fact that the number of hidden layers is directly linked to the complexity, training time, and performance of the model. In this sense, experimental findings suggest that ANN models with two hidden layers are inclined to exhibit more imprecise performance or worse overall results compared to those with a single hidden layer [74,85].

Excessively augmenting the number of neurons inside the hidden layer could deteriorate the learning process, resulting in less accurate models [74]. Similar to having a large number of LVs in the PLS model, which can result in overfitting, a high number of neurons in ANN can also lead to an overfitted model, which loses its generality and predictive ability. For this, during the validation procedure, a subset of external samples is used to assess the performance of ANN models with different numbers of nodes in the hidden layer via the root mean square error of validation (RMSEV). The optimal selection of nodes corresponds to the ANNs with the lowest RMSEV.

Regularization term

A way to avoid overfitting in ANN models is to include a regularization term in the loss function. Two common types are L1 (Lasso regularization) and L2 (Ridge regularization). This latter form is widely adopted and recognized as the most efficient way to prevent overfitting. This term penalizes the sum of the squared amplitude of the model weights. The loss function adopted in this thesis is given by:

$$MSE = MSE + \frac{1}{2}\lambda \sum_{i=1}^{n} w_i^2 \qquad \text{(II.29)}$$

where MSE is the mean squared error, λ is the regularization parameter, and *n* is the total number of weights w_i.

Common values for the regularization parameter typically follow a logarithmic scale ranging from 0 to 0.1, $\lambda \epsilon [0.0001, 0.001, 0.01, 0.1, \ldots]$. Although the value of λ does not alter other well-tuned parameters in the model, its effect is evident in the convergence of the loss function.

Finally, the top-performing network is chosen to advance to the next step, where its performance is evaluated on a test set [86]. Essentially, the ANN undergoes a cycle of training, validation, and subsequent testing steps.

II.4.2.3 Testing step

The test set should consist of samples not included in the training and validation set, ensuring the proper evaluation of the model [87]. To do this, the model accuracy is measured by some loss functions or objective functions such as the root mean square error (RMSE), the determination coefficient (R^2), mean absolute percentage error (MAPE), or the mean absolute error (MAE).

If there is a significant lack of fit, the model should undergo re-training. This involves modifying the learning procedure by incorporating larger datasets that allow for greater accuracy in the results or redefining the model architecture.

There is not a one-size-fits-all solution that ensures the best performance in ANN assays. The right way to develop a high-quality ANN model is to test several architectures, loss functions, and assess predictive ability on validation samples. Only after evaluating several models, a more relevant network can be selected to solve the specific problem at hand.

II.4.3 Autoassociative neural networks

An auto-associative neural network, also known as a replicator neural network or autoencoder, is a feedforward multilayer perceptron network designed to reproduce the input [68,88]. A key feature is that one hidden layer has fewer neurons than the input layer (Figure II-13) which gives this network the aspect of having a bottleneck. The left-hand side of the network, from the input layer to the bottleneck layer, performs a nonlinear mapping of the input $\mathbf{x}$ (e.g., a spectrum) to a reduced latent space so that $\mathbf{x}$ is represented in fewer dimensions. This block is known as the encoder. The right-hand side of the network, from the bottleneck to the output layer, is the decoder part and uses the compressed representation emerging from the bottleneck to approximately reconstruct the input $\mathbf{x}$. The network is trained to reconstruct the input data by minimizing the loss function

$$E = \frac{1}{2}\sum_{i=1}^{I} Q_i \qquad \text{(II.30)}$$

where I is the number of training samples and Q_i is the sum of squared spectral residuals of the i^{th} training sample given by [89]

$$Q_i = \sum_{k=1}^{K} (x_{ik} - \hat{x}_{ik})^2 \qquad \text{(II.31)}$$

where $\mathbf{x}_i = [x_{i1}, \dots x_{ik}, \dots x_{iK}]^{\mathrm{T}}$ is the spectrum of the i^{th} training sample, K is the number of spectral variables and $\hat{\mathbf{x}}_i = [\hat{x}_{i1}, \dots \hat{x}_{ik}, \dots \hat{x}_{iK}]^{\mathrm{T}}$ is the output of the autoencoder.

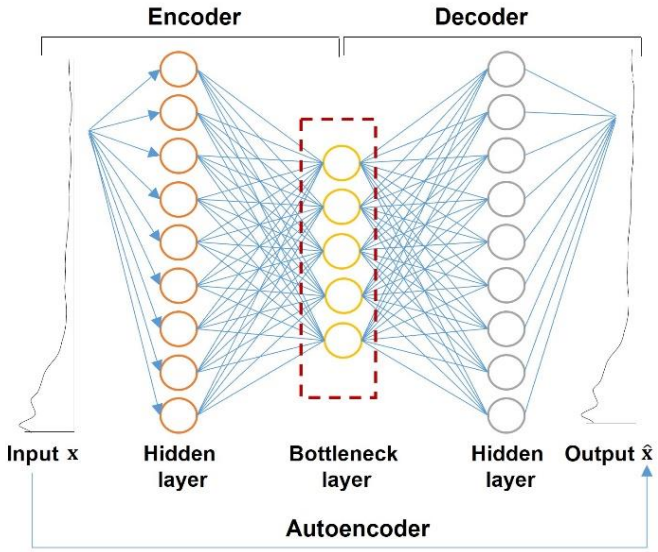


Figure II-13. Autoencoder with three hidden layers. The central layer is the bottleneck layer.

With adequate settings, autoencoders can perform, for example, nonlinear principal component analysis [90]. Although autoencoders are usually symmetric [68], symmetry is not strictly necessary, and the encoder and decoder parts may have a different number of layers and nodes.

II.5 References

[1] D. Skoog, F.J. Holler, S.R. Crouch, Principios de análisis instrumental, 6th ed., Cengage Learning, México, 2008.

[2] D.A. Burns, E.W. Ciurczak, Handbook of Near-Infrared Analysis, 3rd ed., CRC Press Taylor & Francis Group, Boca Raton, Florida, 2008.

[3] C. Pasquini, Near infrared spectroscopy: A mature analytical technique with new perspectives – A review, Anal. Chim. Acta. 1026 (2018) 8–36. https://doi.org/10.1016/j.aca.2018.04.004.

[4] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, N. Jent, A review

of near infrared spectroscopy and chemometrics in pharmaceutical technologies, J. Pharm. Biomed. Anal. 44 (2007) 683–700. https://doi.org/10.1016/j.jpba.2007.03.023.

[5] C.S. Silva, A. Braz, M.F. Pimentel, Vibrational spectroscopy and chemometrics in forensic chemistry: Critical review, current trends and challenges, J. Braz. Chem. Soc. 30 (2019) 2259–2290. https://doi.org/10.21577/0103-5053.20190140.

[6] M.K. Moro, F.D. dos Santos, G.S. Folli, W. Romão, P.R. Filgueiras, A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy, Fuel. 303 (2021) 121283. https://doi.org/10.1016/j.fuel.2021.121283.

[7] I.T. Jolliffe, J. Cadima, Principal component analysis: A review and recent developments, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2016) 20150202. https://doi.org/10.1098/rsta.2015.0202.

[8] D. Granato, J.S. Santos, G.B. Escher, B.L. Ferreira, R.M. Maggio, Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective, Trends Food Sci. Technol. 72 (2018) 83–90. https://doi.org/10.1016/j.tifs.2017.12.006.

[9] L.C. Lee, A.A. Jemain, On overview of PCA application strategy in processing high dimensionality forensic data, Microchem. J. 169 (2021) 106608. https://doi.org/10.1016/j.microc.2021.106608.

[10] I.T. Jolliffe, Principal component analysis: A beginner's guide — II. Pitfalls, myths and extensions, Weather. 48 (1993) 246–253. https://doi.org/10.1002/j.1477-8696.1993.tb05899.x.

[11] R.G. Brereton, Applied Chemometrics for Scientists, John Wiley & Sons Ltd, Chichester (England), 2007. http://www.sciencemediacentre.org/working-with-us/for-scientists/.

[12] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. https://doi.org/10.1039/c3ay41907j.

[13] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 433–459. https://doi.org/10.1002/wics.101.

[14] I.T. Jolliffe, Principal Component Analysis, 2nd ed., Springer Series in Statistics, 2007. https://doi.org/10.1002/ibd.21544.

[15] S. Wang, S. Liu, Y. Yuan, J. Zhang, Z. Wang, X. Che, A novel CC-tSNE-SVR model for rapid determination of diesel fuel quality by near infrared spectroscopy, Infrared Phys. Technol. 106 (2020) 103276. https://doi.org/10.1016/j.infrared.2020.103276.

[16] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[17] H.O. Wold, Soft modelling: the basic design and some extensions, in: Joreskog, K.G. Wold, H.O.A., Systems under Indirect Observations: Part II, North-Holland, Amsterdam, 1982: pp. 1–54.

[18] S. Wold, H. Martens, H. Wold, The multivariate calibration problem in chemistry solved by the PLS method, in: Kågström, B. Ruhe, A., Matrix Pencils, Springer Berlin Heidelberg, 2006: pp. 286--293. https://doi.org/https://doi.org/10.1007/BFb0062108.

[19] P. Geladi, B.R. Kowalski, Partial Least Squares Regression: A Tutorial, Anal. Chim. Acta. 185 (1986) 1–17. https://doi.org/10.1016/0003-2670(86)80028-9.

[20] P. Geladi, H. Martens, L. Hadjiiski, P. Hopke, A calibration tutorial for spectral data. Part 2. Partial least squares regression using Matlab and some neural network results, J. Near Infrared Spectrosc. 4 (1996) 243–255. https://doi.org/10.1255/jnirs.94.

[21] J.M. Andrade-Garda, R. Boqué-Martí, J. Ferré-Baldrich, A. Carlosena-Zubieta, Partial Least-Squares Regression, in: J. Andrade-Garda (Ed.), Basic Chemom. Tech. At. Spectrosc., 2nd ed., The Royal Society of Chemistry, 2009: pp. 187–243. https://doi.org/10.1039/9781847559661-00181.

[22] C.L.M. Morais, M.C.D. Santos, K.M.G. Lima, F.L. Martin, Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach, Bioinformatics. 35 (2019) 5257–5263. https://doi.org/10.1093/bioinformatics/btz421.

[23] X. Xu, T. Liang, J. Zhu, D. Zheng, T. Sun, Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing. 328 (2019) 5–15. https://doi.org/10.1016/j.neucom.2018.02.100.

[24] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148.

https://doi.org/10.1080/00401706.1969.10490666.

[25] Y. Wang, D.J. Veltkamp, B.R. Kowalski, Multivariate Instrument Standardization, Anal. Chem. 63 (1991) 2750–2756. https://doi.org/10.1021/ac00023a016.

[26] W. Yang, Y. Xiong, H. Wang, T. Wu, Y. Du, Interval interaction moving window partial least squares for wavelength interval selection in near infrared spectroscopy, Chemom. Intell. Lab. Syst. 241 (2023) 104976. https://doi.org/10.1016/j.chemolab.2023.104976.

[27] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32. https://doi.org/10.1016/j.aca.2010.03.048.

[28] R.M. Balabin, S. V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72. https://doi.org/10.1016/j.aca.2011.03.006.

[29] R. Leardl, L. Norgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497. https://doi.org/10.1002/cem.893.

[30] J.H. Jiang, R. James, B.H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565. https://doi.org/10.1021/ac011177u.

[31] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta. 501 (2004) 183–191. https://doi.org/10.1016/j.aca.2003.09.041.

[32] C. Ranzan, L.F. Trierweiler, B. Hitzmann, J.O. Trierweiler, NIR pre-selection data using modified changeable size moving window partial least squares and pure spectral chemometrical modeling with ant colony optimization for wheat flour characterization, Chemom. Intell. Lab. Syst. 142 (2015) 78–86. https://doi.org/10.1016/j.chemolab.2015.01.007.

[33] S. Wiklund, D. Nilsson, L. Eriksson, M. Sjöström, S. Wold, K. Faber, A randomization test for PLS component selection, J. Chemom. 21 (2007) 427–439. https://doi.org/10.1002/cem.1086.

[34] T. Tran, E. Szymańska, J. Gerretzen, L. Buydens, N.L. Afanador, L. Blanchet,

Weight randomization test for the selection of the number of components in PLS models, J. Chemom. 31 (2017) 1–15. https://doi.org/10.1002/cem.2887.

[35] 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, 88 (2007) 183–188. https://doi.org/10.1016/j.chemolab.2007.04.006.

[36] C.L. Cunha, A.S. Luna, R.C.G. Oliveira, G.M. Xavier, M.L.L. Paredes, A.R. Torres, Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate calibration, 204 (2017) 185–194. https://doi.org/10.1016/j.fuel.2017.05.057.

[37] M.G. Nespeca, R.R. Hatanaka, D.L. Flumignan, E. De Oliveira, Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis, J. Anal. Methods Chem. (2018) 1795624. https://doi.org/10.1155/2018/1795624.

[38] G.E. Fodor, R.A. Mason, S.A. Hutzler, Estimation of middle distillate fuel properties by FT-IR, Appl. Spectrosc. 53 (1999) 1292–1298. https://doi.org/10.1366/0003702991945542.

[39] V.O. Santos Jr., 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. 547 (2005) 188–196. https://doi.org/10.1016/j.aca.2005.05.042.

[40] L. de F. Bezerra de Lira, F.V. Cruz de Vasconcelos, C. Fernandes Pereira, A.P. Silveira Paim, L. Stragevitch, M.F. Pimentel, Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration, Fuel. 89 (2010) 405–409. https://doi.org/10.1016/j.fuel.2009.05.028.

[41] J.C.L. Alves, C.B. Henriques, R.J. Poppi, Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system, Fuel. 97 (2012) 710–717. https://doi.org/10.1016/j.fuel.2012.03.016.

[42] American Society for Testing Materials, ASTM E1655-17 Standard Practices for Infrared Multivariate Quantitative Analysis, (2017) 1–29. https://doi.org/10.1520/E1655-17.

[43] S.K. Setarehdan, J.J. Soraghan, D. Littlejohn, D.A. Sadler, Maintenance of a calibration model for near infrared spectrometry by a combined principal

component analysis-partial least squares approach, Anal. Chim. Acta. 452 (2002) 35–45. https://doi.org/10.1016/S0003-2670(01)01446-5.

[44] T. Næs, H. Martens, Multivariate calibration. II. Chemometric methods, Trends Anal. Chem. 3 (1984) 266–271. https://doi.org/10.1016/0165-9936(84)80044-8.

[45] H. Hotelling, Multivariate quality control, illustrated by the air testing of sample bombsights, in: C. Eisenhart, M.W. Hast. W.A. Wallis (Eds.), Tech. Stat. Anal., McGraw-Hill, New York, 1947: pp. 113–184.

[46] J.A. Westerhuis, S.P. Gurden, A.K. Smilde, Standardized Q-statistic for improved sensitivity in the monitoring of residuals in MSPC, J. Chemom. 14 (2000) 335–349. https://doi.org/10.1002/1099-128X(200007/08)14:4<335::AID-CEM579>3.0.CO;2-F.

[47] B.M. Wise, R.T. Roginski, A calibration model maintenance roadmap, IFAC-PapersOnLine. 28 (2015) 260–265. https://doi.org/10.1016/j.ifacol.2015.08.191.

[48] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Transfer of Multivariate Calibration Models, Chemom. Intell. Lab. Syst. 64 (2002) 181–192. https://doi.org/10.1016/B978-044452701-1.00077-6.

[49] C.F. Pereira, M.F. Pimentel, R.K.H. Galvão, F.A. Honorato, L. Stragevitch, M.N. Martins, A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers, Anal. Chim. Acta. 611 (2008) 41–47. https://doi.org/10.1016/j.aca.2008.01.071.

[50] J.J. Workman, A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy, Appl. Spectrosc. 72 (2018) 340–365. https://doi.org/10.1177/0003702817736064.

[51] T. Fearn, Standardisation and Calibration Transfer for near Infrared Instruments: A Review, J. Near Infrared Spectrosc. 9 (2001) 229–244. https://doi.org/10.1255/jnirs.309.

[52] B. Malli, A. Birlutiu, T. Natschläger, Standard-free calibration transfer - An evaluation of different techniques, Chemom. Intell. Lab. Syst. 161 (2017) 49–60. https://doi.org/10.1016/j.chemolab.2016.12.008.

[53] P. Mishra, R. Nikzad-Langerodi, F. Marini, J.M. Roger, A. Biancolillo, D.N. Rutledge, S. Lohumi, Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always, TrAC - Trends Anal. Chem. 143 (2021) 116331. https://doi.org/10.1016/j.trac.2021.116331.

[54] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Networks. 22 (2011) 199–210. https://doi.org/10.1109/TNN.2010.2091281.

[55] E. Andries, Penalized eigendecompositions: motivations from domain adaptation for calibration transfer, J. Chemom. 31 (2017) 1–14. https://doi.org/10.1002/cem.2818.

[56] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, Two standard-free approaches to correct for external influences on near-infrared spectra to make models widely applicable, Postharvest Biol. Technol. 170 (2020) 111326. https://doi.org/10.1016/j.postharvbio.2020.111326.

[57] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, FRUITNIR-GUI: A graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction, Postharvest Biol. Technol. 175 (2021) 111414. https://doi.org/10.1016/j.postharvbio.2020.111414.

[58] P. Mishra, CT-GUI: A graphical user interface to perform calibration transfer for multivariate calibrations, Chemom. Intell. Lab. Syst. 214 (2021) 104338. https://doi.org/10.1016/j.chemolab.2021.104338.

[59] M. Ghifary, D. Balduzzi, W.B. Kleijn, M. Zhang, Scatter component analysis: A unified framework for domain adaptation and domain generalization, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1414–1430. https://doi.org/10.1109/TPAMI.2016.2599532.

[60] R. Nikzad-Langerodi, W. Zellinger, E. Lughofer, S. Saminger-Platz, Domain-Invariant Partial-Least-Squares Regression, Anal. Chem. 90 (2018) 6693–6701. https://doi.org/10.1021/acs.analchem.8b00498.

[61] J. Zhang, B. Li, Y. Hu, L. Zhou, G. Wang, G. Guo, Q. Zhang, S. Lei, A. Zhang, A parameter-free framework for calibration enhancement of near-infrared spectroscopy based on correlation constraint, Anal. Chim. Acta. 1142 (2021) 169–178. https://doi.org/10.1016/j.aca.2020.11.006.

[62] M. Zeaiter, J.M. Roger, V. Bellon-Maurel, Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations, Chemom. Intell. Lab. Syst. 80 (2006) 227–235. https://doi.org/10.1016/j.chemolab.2005.06.011.

[63] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, B. Moser, Domain-invariant regression under beer-lambert's law, Proc. - 18th IEEE Int. Conf. Mach. Learn.

Appl. ICMLA 2019. (2019) 581–586. https://doi.org/10.1109/ICMLA.2019.00108.

[64] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, B.A. Moser, Domain adaptation for regression under Beer–Lambert's law, Knowledge-Based Syst. 210 (2020) 106447. https://doi.org/10.1016/j.knosys.2020.106447.

[65] P. Mishra, R. Nikzad-Langerodi, Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit, Infrared Phys. Technol. 111 (2020) 103547. https://doi.org/10.1016/j.infrared.2020.103547.

[66] B. Mikulaseka, V. Fonseca Diaz, C. Herwig, R. Nikzad-Langerodi, Partial Least Squares Regression With Multiple Domains, (2022). https://doi.org/10.13140/RG.2.2.23750.75845.

[67] X. Capron, B. Walczak, O.E. De Noord, D.L. Massart, Selection and weighting of samples in multivariate regression model updating, Chemom. Intell. Lab. Syst. 76 (2005) 205–214. https://doi.org/10.1016/j.chemolab.2004.11.003.

[68] C.C. Aggarwal, Neural Networks and Deep Learning, Springer, Yorktown Heights, New York, 2018. https://doi.org/10.1007/978-3-031-03758-0_5.

[69] R.M. Balabin, E.I. Lomakina, R.Z. Safieva, Neural network (ANN) approach to biodiesel analysis : Analysis of biodiesel density , kinematic viscosity , methanol and water contents using near infrared (NIR) spectroscopy, Fuel. 90 (2010) 2007–2015. https://doi.org/10.1016/j.fuel.2010.11.038.

[70] H.A.G. Al-kaf, K.S. Chia, N.A.M. Alduais, A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum, Pet. Sci. Technol. 36 (2018) 411–418. https://doi.org/10.1080/10916466.2018.1425717.

[71] J.S. Oliveira, R. Montalvão, L. Daher, P.A.Z. Suarez, J.C. Rubim, Determination of methyl ester contents in biodiesel blends by FTIR-ATR and FTNIR spectroscopies, Talanta. 69 (2006) 1278–1284. https://doi.org/10.1016/j.talanta.2006.01.002.

[72] B. Basu, G.S. Kapur, A.S. Sarpal, R. Meusinger, A Neural Network Approach to the Prediction of Cetane Number of Diesel Fuels Using Nuclear Magnetic Resonance (NMR) Spectroscopy, Energy and Fuels. 17 (2003) 1570–1575. https://doi.org/10.1021/ef030083f.

[73] M. Tkáč, R. Verner, Artificial neural networks in business: Two decades of research, Appl. Soft Comput. J. 38 (2016) 788–804.

https://doi.org/10.1016/j.asoc.2015.09.040.

[74] B. Karlik, A.V. Olgac, Performance Analysis of Various Activation Functions in Artificial Neural Networks, J. Artif. Intell. Expert Syst. 1 (2012) 111–122. https://doi.org/10.1088/1742-6596/1237/2/022030.

[75] V.. V. Phansalkar, P.S. Sastry, Analysis of Back-Propagation Algorithm with Momentum, IEEE Trans. Neural Networks. 5 (1994) 505–506.

[76] F. Cao, H. Ye, D. Wang, A probabilistic learning algorithm for robust modeling using neural networks with random weights, Inf. Sci. (Ny). 313 (2015) 62–78. https://doi.org/10.1016/j.ins.2015.03.039.

[77] T. Hastie, R. Tibshirani, J. Friedman, The elements of Statistical Learning. Data Mining, Inference, and Prediction., Springer Series in Statistics, Stanford, California August 2008, 2009.

[78] A.B. Badiru, D.B. Sieger, Neural network as a simulation metamodel in economic analysis of risky projects, Eur. J. Oper. Res. 105 (1998) 130–142. https://doi.org/10.1016/S0377-2217(97)00029-5.

[79] J. Sola, J. Sevilla, Importance of input data normalization for the application of neural networks to complex industrial problems, IEEE Trans. Nucl. Sci. 44 (1997) 1464–1468. https://doi.org/10.1109/23.589532.

[80] I. Kaastraa, M. Boydb, Designing a neural network for forecasting financial time series 29, Neurocomputing. 10 (1996) 215–236. https://doi.org/10.1016/0925-2312(95)00039-9.

[81] S. Dündar, I. Şahin, Train re-scheduling with genetic algorithms and artificial neural networks for single-track railways, Transp. Res. Part C Emerg. Technol. 27 (2013) 1–15. https://doi.org/10.1016/j.trc.2012.11.001.

[82] J. Zhang, A.J. Morris, A sequential learning approach for single hidden layer neural networks, Neural Networks. 11 (1998) 65–80. https://doi.org/10.1016/S0893-6080(97)00111-1.

[83] V. Martínez-Martínez, F.J. Gomez-Gil, J. Gomez-Gil, R. Ruiz-Gonzalez, An Artificial Neural Network based expert system fitted with Genetic Algorithms for detecting the status of several rotary components in agro-industrial machines using a single vibration signal, Expert Syst. Appl. 42 (2015) 6433–6441. https://doi.org/10.1016/j.eswa.2015.04.018.

[84] J. Heaton, Introduction to neural networks with Java, Heaton Research, Inc.,

United States of America, 2008.

[85] J. de Villiers, E. Barnard, Backpropagation Neural Nets with One and Two Hidden Layers, IEEE Trans. Neural Networks. 4 (1993) 136–141. https://doi.org/10.1109/72.182704.

[86] J. Zhang, E.B. Martin, A.J. Morris, C. Kiparissides, Inferential estimation of polymer quality using stacked neural networks, Comput. Chem. Eng. 21 (1997) 1025–1030. https://doi.org/10.1016/s0098-1354(97)87637-5.

[87] F.M. De Oliveira, L.S. De Carvalho, L.S.G. Teixeira, C.H. Fontes, K.M.G. Lima, A.B.F. Câmara, H.O.M. Araújo, R. V. Sales, Predicting Cetane Index, Flash Point, and Content Sulfur of Diesel-Biodiesel Blend Using an Artificial Neural Network Model, Energy and Fuels. 31 (2017) 3913–3920. https://doi.org/10.1021/acs.energyfuels.7b00282.

[88] J. Schmidhuber, Deep Learning in neural networks: An overview, Neural Networks. 61 (2015) 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

[89] S. Hawkins, H. He, G. Williams, R. Baxter, Outlier detection using replicator neural networks, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2454 LNCS (2002) 170–180. https://doi.org/10.1007/3-540-46145-0_17.

[90] M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE J. 37 (1991) 233–243. https://doi.org/10.1002/aic.690370209.

Chapter III
Exploration of properties and spectra

III.1 Description of the diesel samples

The diesel samples used in this study are 2307 samples from 36 months of production between 2018 and 2021. These samples correspond to desulfurized diesel (stream 1) and commercial diesel (stream 2), the former being the main constituent in the blend used to formulate commercial diesel. In this work, the temperature at 95% recovered, flash point, cloud point, density, cetane number, and sulfur content were determined for desulfurized diesel. These properties were also determined for commercial diesel, together with the temperature at 65% and 85% recovered, viscosity, cold filter plugging point (CFPP), and FAME content. The collected data correspond to analyses carried out following production requirements, so there are more commercial diesel samples than desulfurized samples (Table III-1). Not all the samples had the same properties determined. For this reason, the number of samples for each property ranged between 75 and 1583.

Table III-1. Number of samples for each property.

Property	Desulfurized diesel	Commercial diesel
T95%	147	1581
Flash point	146	1583
Cloud point	90	1006
Density	150	1578
Cetane number	75	225
Sulfur content	146	1550
T65%	-	1578
T85%	-	1582
Viscosity	-	229
Cold filter plugging point (CFPP)	-	1573
FAME content	-	979

III.2 Distribution of properties in the samples of each stream

Figure III-1 compares the distribution of the properties for the two types of diesel. Although the ranges of property values for both streams have some overlap, the values of T95% recovered, flash point, cloud point, density, cetane number, and sulfur content of desulfurized diesel tend to be higher than for commercial diesel.

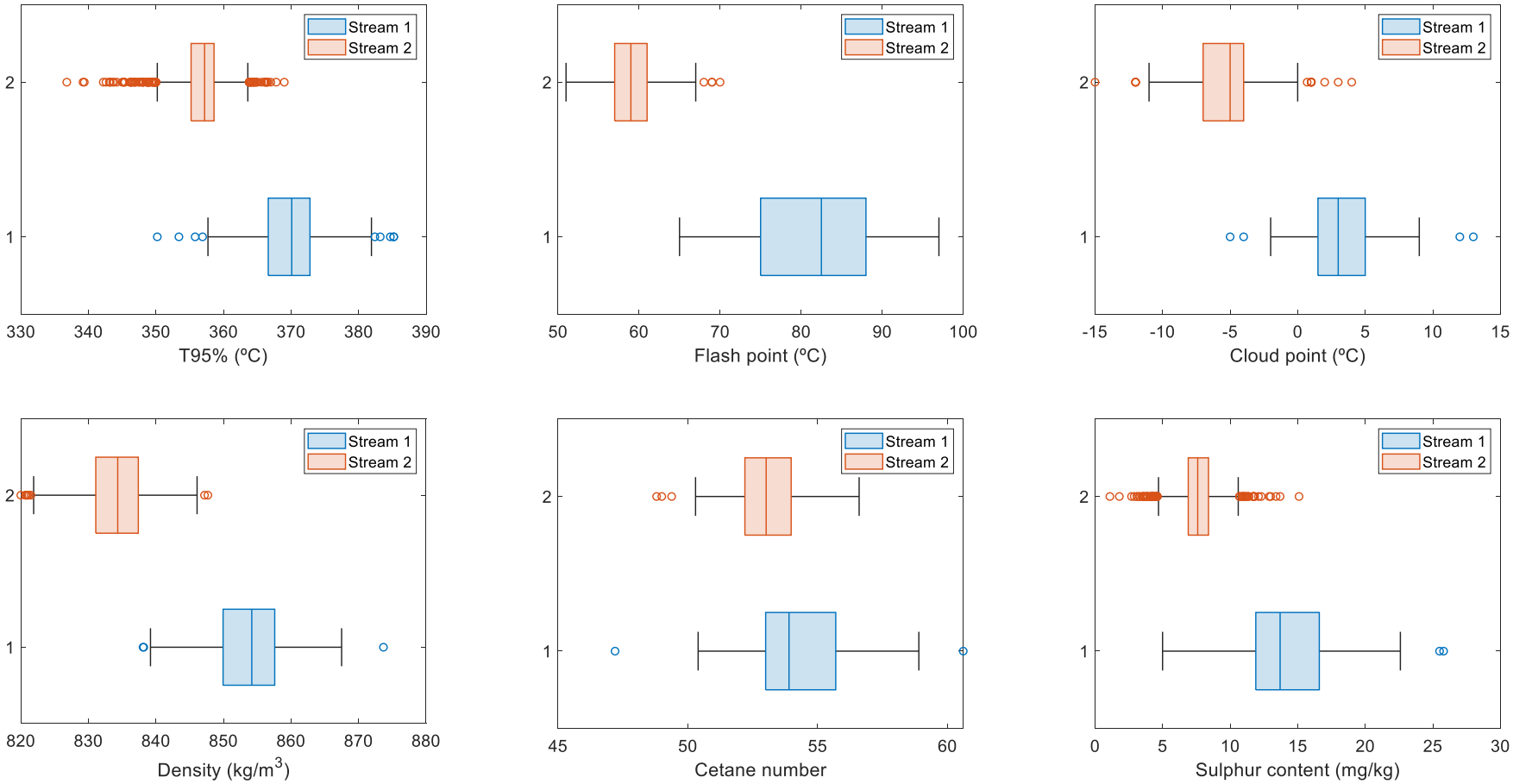


Figure III-1. Box chart of the measured diesel properties from streams 1 and 2.

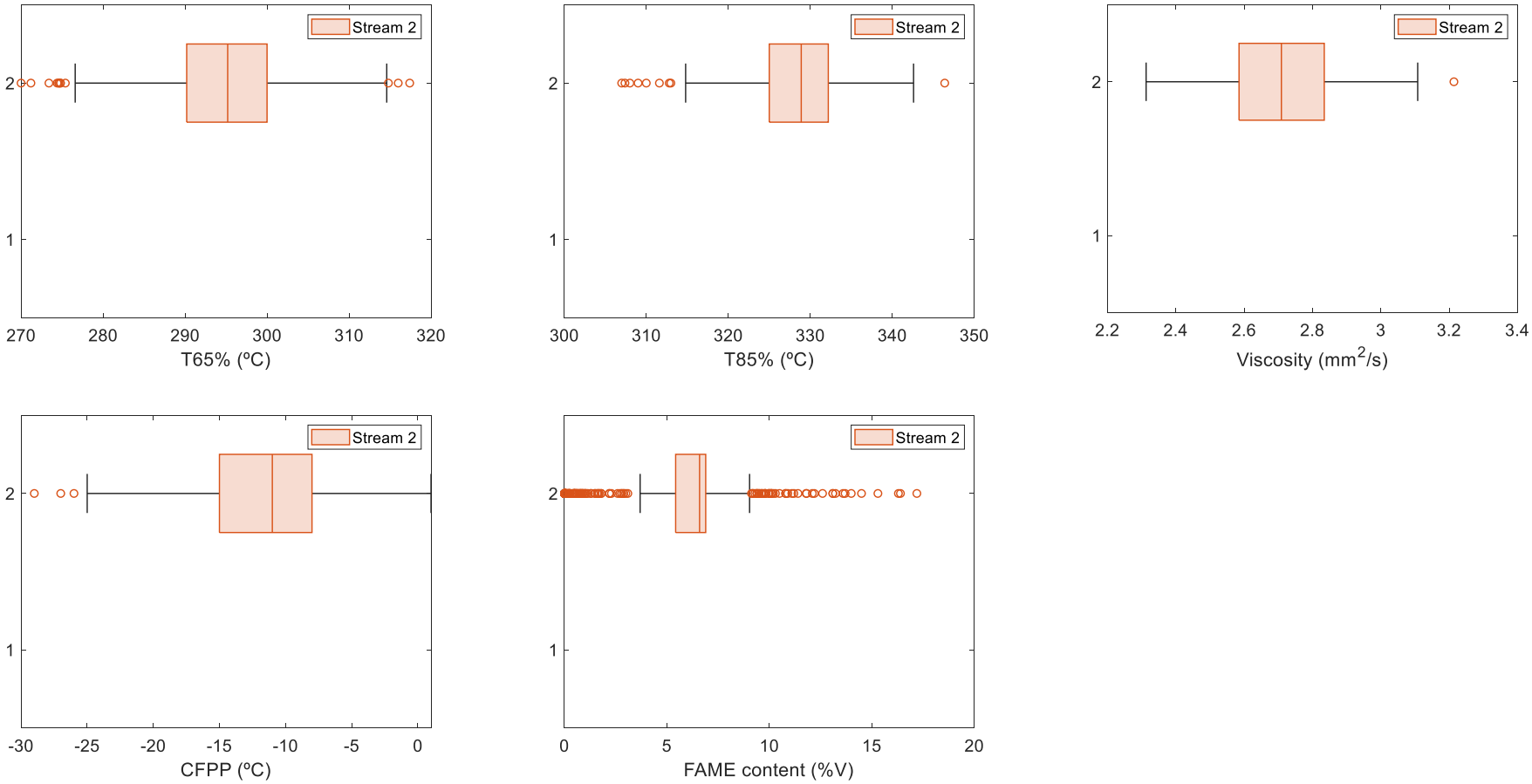


Figure III-1 (cont.). Box chart of the measured diesel properties from streams 1 and 2.

III.3 Correlation between physicochemical properties

Figure III-2 and Figure III-3 show the heat map of correlation coefficients (r) among properties pairs in streams 1 and 2, respectively.

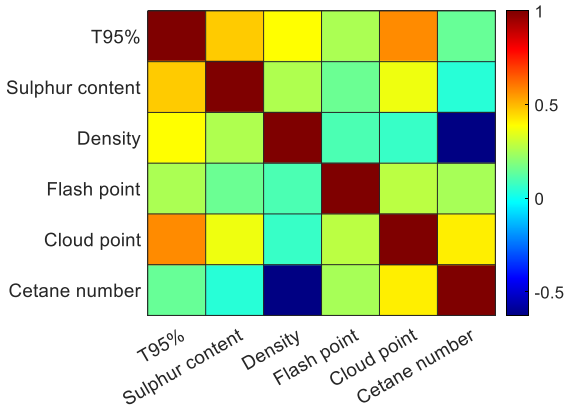


Figure III-2. Heat map of correlation coefficients between all pairs of properties in stream 1.

As can be seen, correlations are weak between the properties of stream 1, with correlation coefficients ranging between -0.626 and 0.568. The most positively correlated pair was T95% and density (r = 0.568). This correlation can be expected if one considers that both density and T95% recovered are measures of the amount of heavy fractions (naphthenes and aromatics) in diesel, meaning that higher density samples may contain heavier fractions of hydrocarbons and this will result in higher distillation temperatures [1]. In contrast, lower-density samples may contain lighter fractions of hydrocarbons (paraffin) and, thus, lower distillation temperatures, especially in the first fractions of the distillation. Both density and distillation temperatures (in this case, T95% recovered) increase with the number of carbons for compounds of the same hydrocarbon class and in the following order: isoparaffins < n-paraffins < naphthenes < aromatics for compounds with the same carbon number [2,3]. The most negative correlation is between density and cetane (r = -0.626). This can be associated with the fact that cetane number is defined from the relative proportions of *n*-hexadecane or cetane (CN = 100 arbitrarily assigned) and *α*-methylnaphthalene (CN = 0 also arbitrarily assigned), which have the same ignition retardation properties as the test fuel. Thus, higher-density samples may contain heavier hydrocarbons, which could affect ignition characteristics and, consequently, the cetane number. CN increases in the following order: aromatics < isoparaffins and naphthenes < paraffin. Last, the least correlated pairs were cetane number and sulfur content (r = 0.045) and cloud point and density (r = 0.068).

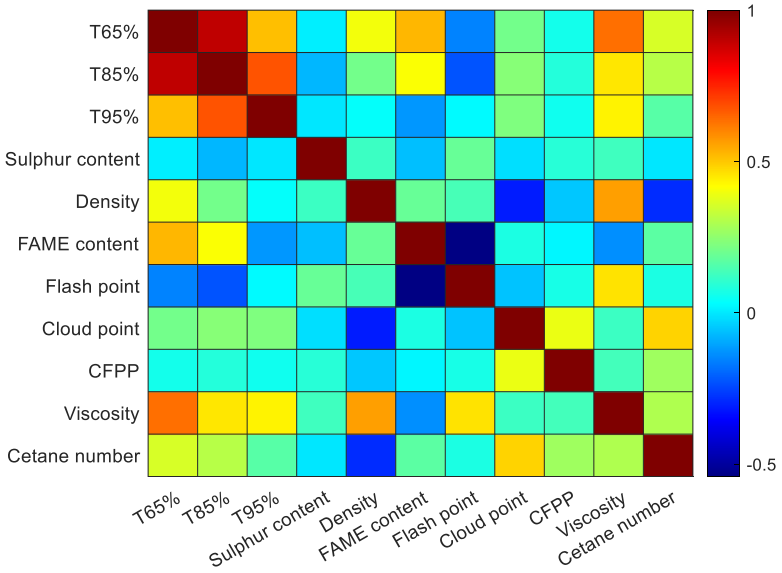


Figure III-3. Heat map of correlation coefficients between all pairs of properties in stream 2.

In contrast, some correlations were higher between the properties of stream 2, with correlation coefficients ranging between -0.541 and 0.90. The most positively correlated pairs were: T65% and T85% (r = 0.901), T85% and T95% (r = 0.876), and to a lesser extent T65% and T95% (r = 0.517). As expected, the distillation temperatures at 65%, 85%, and 95% recovered, which are the most important parameters of the distillation curve [4], are typically highly correlated (see Figure III-4). The composition, including the class and structures of hydrocarbons present, influences the boiling points and, consequently, the distillation profile of the diesel fuel. The higher the distillation temperatures in the final section of the distillation curve (in this case, T65%, T85%, and T95% recovered), the higher the heavier hydrocarbon fractions in the diesel.

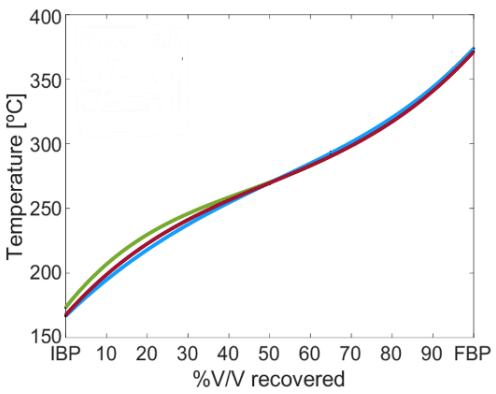


Figure III-4. Distillation curves of three diesel samples.

Furthermore, the pairs T65% and viscosity (r = 0.639) and between density and viscosity (r = 0.561) were also positively correlated. These correlations can be expected if one considers that viscosity is a measure of how easily diesel fuel can flow and is related to the molecular weight of the compounds rather than to the class of hydrocarbons [2]. Naphthenes generally exhibit slightly higher viscosity than paraffins and aromatics for a given number of carbons. It is unsurprising that the correlation between viscosity and distillation temperatures decreases as the percentage of recovered volume increases. As the distillation process progresses, the composition of the fractions that remain to be distillate diverges further from the original diesel composition. A similar trend for the correlation between density and distillation temperatures can also be deduced from the heat map in Figure III-3.

Finally, as expected, FAME and cetane number (r = 0.525) were also positively correlated. In this sense, the higher the FAME content, the better the ignition properties and, therefore the higher the cetane number of the fuel. Similar to the aforementioned properties, cetane number has a systematic variation with hydrocarbon class (paraffins, isoparaffins, naphthene, or aromatic) and hydrocarbon structure (linear or branched). Within the characteristic range of each class, the higher the molecular weight in paraffins and longer side chains in isoparaffins, in high molecular weight naphthenes, and in single-ring aromatics, the higher the cetane number. In this sense, FAME predominantly consists of long-chain hydrocarbon groups characterized by minimal branching or aromatic structures [8], resulting in a typically higher cetane number (CN).

The most negative correlation is between FAME and flash point (r = -0.541). This observation is intriguing, considering that FAME commonly exhibits higher flash points compared to diesel [5,6]. Besides, it is also known that light compounds show higher flash points (FP) [7]. Therefore, although the reason behind this correlation is not clear, the effect of FAME content is probably masked by the effect of the amount of heavier non-FAME fractions in the samples.

The least correlated pairs were T65% and sulfur content (r = 0.011), T65% and CFPP (r = 0.062), T85% and sulfur content (r = -0.075), T85% and CFPP (r = 0.085), T95% and sulfur content (r = -0.003), T95% and density (r = 0.04), T95% and CFPP (r = 0.061), sulfur content and FAME content (r = -0.064), sulfur content and cloud point (r = -0.017), sulfur content and cetane number (r = -0.002), density and CFPP (r = -0.048), FAME and cloud point (r = 0.074), FAME and CFPP (r = 0.023), flash point and cloud point (r = -0.056), flash point and CFPP (r = 0.072) and flash point and cetane number (r = 0.078).

III.4 Acquisition and spectral analysis

III.4.1 Spectral acquisition

The infrared spectra of the samples were recorded in absorbance mode with an Analect Diamond 20 FTIR/FT-NIR process lab analyzer used for routine analysis under laboratory conditions. The 100% of the transmittance was established daily with the empty cell. Diesel samples were injected into a flow cell of 0.5 mm path length. The flow cell was flushed with toluene between samples. The spectra were recorded at room temperature (*22 ± 2* ºC) under the following conditions: wavenumber range 7000-986 cm^{-1}, resolution 4 cm^{-1}, and averaging of 64 scans. A new background spectrum was acquired daily to keep up with baseline shifts and environmental fluctuations. Figure III-5 shows the spectra of the total set of samples analyzed for each stream (150 and 1583 for stream 1 and stream 2, respectively). It can be seen that the absorbances recorded in the following subregions of the IR spectrum: 3276 - 2755 cm^{-1}, 1750 -1735 $cm^{-1,}$ and 1555 - 1320 cm^{-1} for all samples were very high. This is probably attributable to the heightened molar absorptivities of active species in these mid-infrared (MIR) subregions, which often need sample dilution before analysis. Only the spectral regions with absorbance lower than 1.5 were considered for the classical spectral analysis.

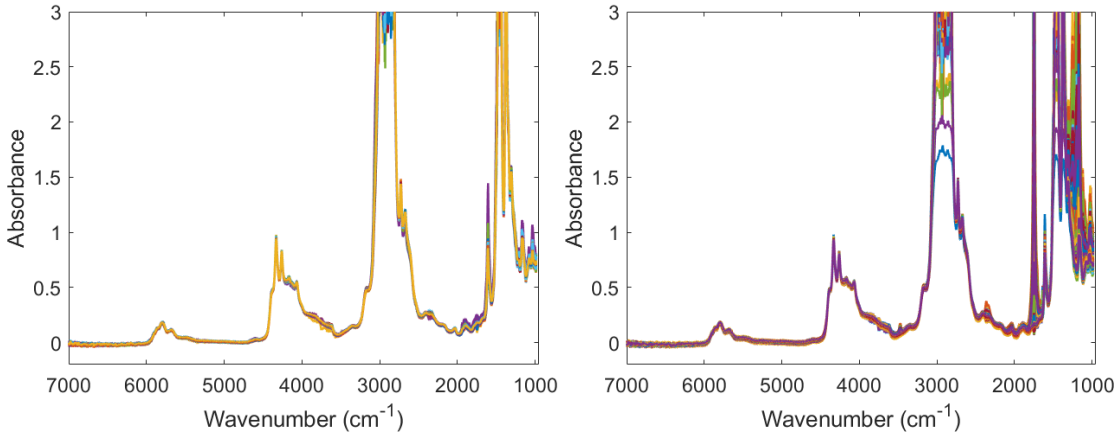


Figure III-5. Infrared spectra of the analyzed samples from stream 1 (left) and stream 2 (right).

III.4.2 Classical spectral analysis

Figure III-6 shows the spectrum of a sample from each stream, both with the same cetane number value (53). For better analysis, the spectrum is divided into three spectral subregions: NIR-MIR (6024.5–3270.7 cm^{-1}), MIR (2753.9–1554.3 cm^{-1}), and MIR (1320 – 984 cm^{-1}).

The NIR absorption bands 6021 and 4000 cm^{-1} (Figure III-6A) are the result of the first overtones and combination regions of the C-H bond stretching, and the second overtone of the C=O and combination regions of C-O of hydrocarbons characteristic of diesel [12–14]. Specifically, the absorption bands of the C-H bond stretching of the methyl group in branched alkanes appear around 5905 cm^{-1}, 5872 cm^{-1}, and 4100 cm^{-1}. Some bands near 5800 cm^{-1}, 5680 cm^{-1}, and 4336 cm^{-1} correspond to the C-H bond stretching of the methylene group in alkanes [13]. The combination band of the C-H and C=O bond appears near 4650 cm^{-1} [12,13]. For some spectra, some differences are observed in the absorption band near 4425 cm^{-1}, probably related to the vibrational mode of stretching of the C-H bond in the terminal methyl group near the carbonyl group.

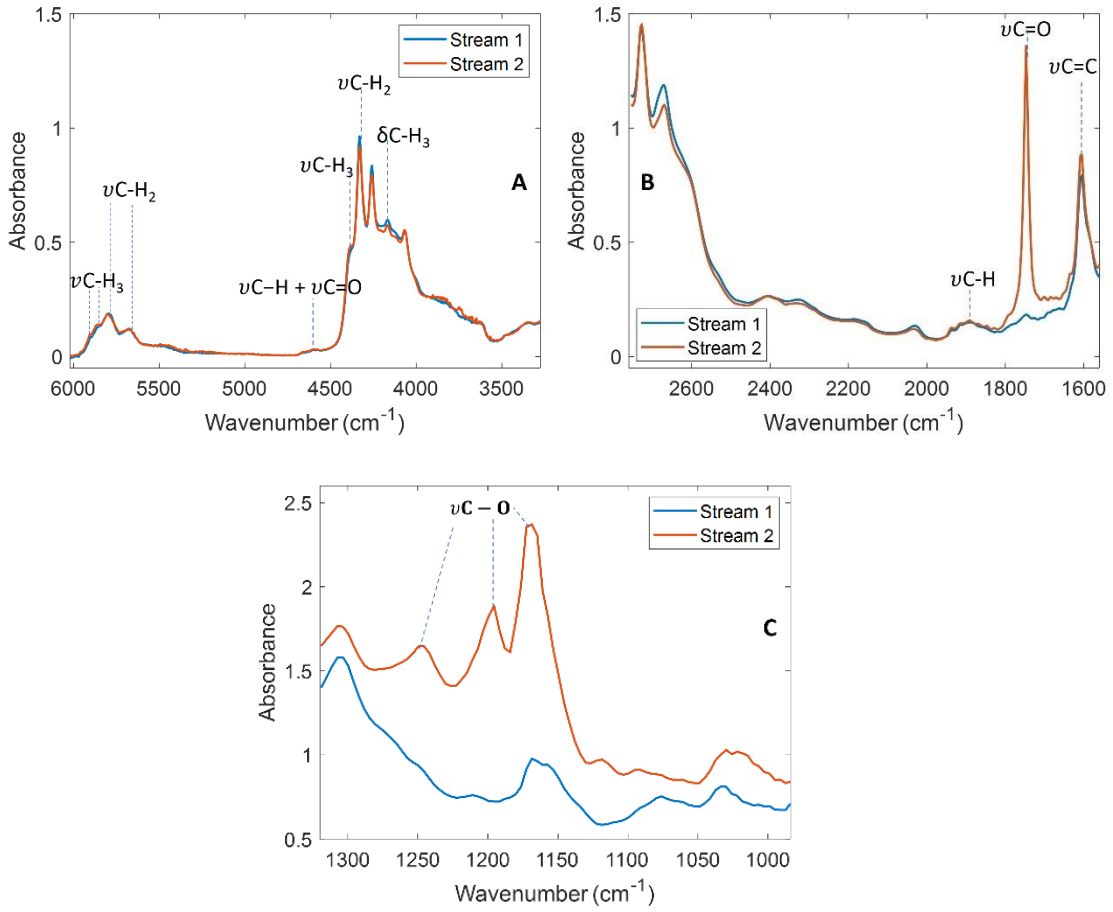


Figure III-6. Spectra of diesel of streams 1 and 2 in the region a) 6021 and 4000 cm^{-1}, b) 2760-1560 cm^{-1} and c) 1320-984 cm^{-1}.

In the MIR region, the absorption bands between 2760 and 984 cm^{-1} (Figure III-6B and Figure III-6C) result from the fundamental vibrational modes of the groups $C=O$ and $C-C$ [15]. The weak bending bands of the $C-H$ bond (overtone) of aromatic compounds

are observed around 2000-1650 cm^{-1}. The content of FAME in the samples produces an intense carbonyl absorption band (1750 and 1735 cm^{-1}), corresponding to the stretching vibrations of the $C=O$ bond, and two aliphatic ester absorption bands (1300 and 1050 cm^{-1}) characteristic of two stretching vibrations of the $C-O$ bond [16,17]. Specifically, the coupled asymmetric vibrations correspond to the bonds ($CC(=O)-O$ and $O-C-C$) around 1250 and 1205 cm^{-1}, respectively [14,18,19]. The band at 1175 cm^{-1} provides additional spectral information and completes the three-band pattern (1250, 1205, and 1175 cm^{-1}) characteristic of the methyl esters of long-chain fatty acids [14,18,19]. The vibrational group attribution to each band is presented in Table III-2.

Table III-2. Vibrational group and spectral bands for diesel samples.

Vibrational group attribution	Wavenumber (cm^{-1})
$\mathbf{CH_3}$ stretch	5905, 5872, 4100
$\mathbf{CH_2}$ stretch	5800, 5680, 4336
$\mathbf{C-H + C=O}$ combination	4650
$\mathbf{CH_3}$ stretch	4425
$\mathbf{C=O}$ carbonyl stretch	1750 - 1735
$\mathbf{C=C}$ stretch (alkenes)	1660-1600
$\mathbf{C=C}$ stretch (aromatic)	1600, 1475
$\mathbf{CC(=O)-O}$ stretch (aliphatic ester)	1250
$\mathbf{O-C-C}$ stretch (aliphatic ester)	1205
$\mathbf{O-C-C}$ stretch (methyl esters of long-chain fatty acids)	1175

III.4.3 Multivariate spectral analysis

Principal component analysis (PCA) was carried out with spectra of samples from each stream, for which reference values for the properties of interest were simultaneously available. A matrix $\mathbf{X} = [\mathbf{X_1}; \mathbf{X_2}](261 \times 963)$ was constructed and mean-centered, where $\mathbf{X_1}$: 74×963 and $\mathbf{X_2}$: 187×963 are the spectral matrices of streams 1 and 2, respectively, in the ranges 6024.5 – 3270.7 cm^{-1} and 2753.9 – 1554.3 cm^{-1}.

The first three principal components explain 44.27%, 27.18%, and 9.43% of the variance, respectively, and 80.88% of the accumulated variance. Figure III-7 shows the PCA scores using the FAME content to evidence the variability between the two types of diesel samples. The graph of PC2 vs. PC1 (Figure III-7A) shows that the spectra of both streams overlap to some degree. Although this overlap is maintained in Figure III-7B, two groups of samples are distinguished. One group consists mainly of samples from

stream 2 with FAME located at the top of Figure III-7B. The other group is formed by the remaining samples from stream 2 (they do not contain FAME or this is less than 0.05% V/V) and the samples from stream 1, located at the bottom of Figure III-7B. This differentiation is achieved by PC3 despite its low percentage of explained variance.

The PCA of the spectra in the aforementioned working regions could not differentiate the samples of stream 1 from the samples of stream 2. However, it seems that PCA was able to differentiate the samples by their FAME content. This behavior was already expected since the absorption band of the C=O bond linked to FAME content is the most significant spectral difference between the samples of the analyzed streams.

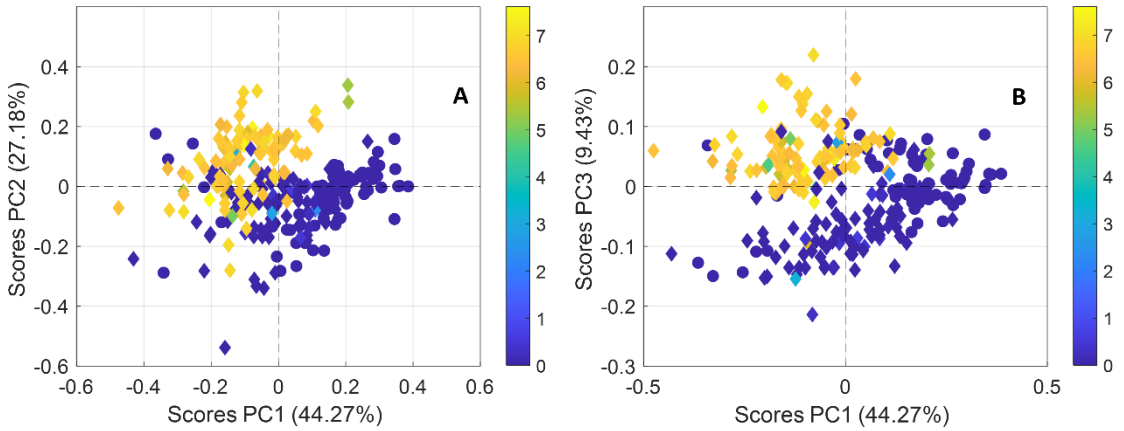


Figure III-7. Score plots of PCA of the spectra of the diesel samples: stream 1 (circles) and stream 2 (rhombus) and the FAME content measured in each sample (colour bar).

PCA may encounter limitations as it consistently seeks linear relationships [20,21]. In order to address this issue, t-SNE from the same spectral data can account for non-linear relationships, resulting in better differentiation between samples from both streams. Figure III-8 depicts the spectral data reduced to two dimensions after applying t-SNE with Euclidean and cosine distance metrics. It can be seen that the first dimension separates the spectra, taking into account their FAME content, and the second dimension differentiates the samples of stream 1 from the samples of stream 2. Three clusters are evident: i) samples from stream 1, ii) samples from stream 2 with FAME content, and iii) samples from stream 2 without FAME content. In addition, one notices that some spectra from stream 1 seem to belong to cluster 3, which is not surprising considering that the samples from stream 1 are the basis of the chemical formulation of samples from stream 2. In this sense, the lower the FAME content, the greater the similarity between the spectra of clusters 1 and 3. Overall, t-SNE provides better

clustering than PCA since the expected differences between the samples are more distinguishable. Besides, t-SNE using cosine distance outperformed Euclidean distance since a better separation between clusters was obtained.

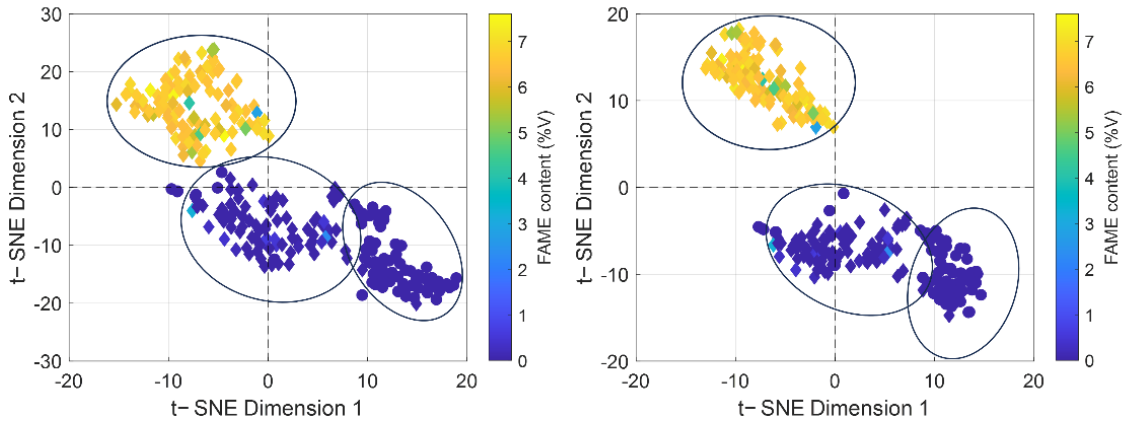


Figure III-8. Plot of the spectral data {stream 1 (circles) and stream 2 (rhombus) and the FAME content measured in each sample (colour bar)} after t-SNE by using euclidean (left) and cosine (right) distance metrics, perplexity = 30, learning rate = 500.

Figure III-9 shows the spectral data reduced to 2 dimensions by applying t-SNE (cosine distance) and the reference values of the six common properties for both streams as code colour. It can be seen that the spectral characteristics of the samples from each stream, which are a reflection of their chemical composition, play a pivotal role in determining their property behavior. Specifically, spectra associated with samples from stream 1 exhibit higher values for density, T95%, flash point, cloud point, cetane number, and sulfur content compared to those from stream 2. Interestingly, among the spectra corresponding to stream 2, despite variations in ester content, no significant differences are observed in their property values. Also note that within the same cluster, remarkably similar spectra are linked to different reference values, and across different clusters, several spectra share the same reference values.

This exploratory analysis suggests that, despite the identification of three groups of samples based on their spectral characteristics, the values of the physicochemical properties are strongly influenced by the origin of the samples (streams 1 or 2) rather than by the FAME content.

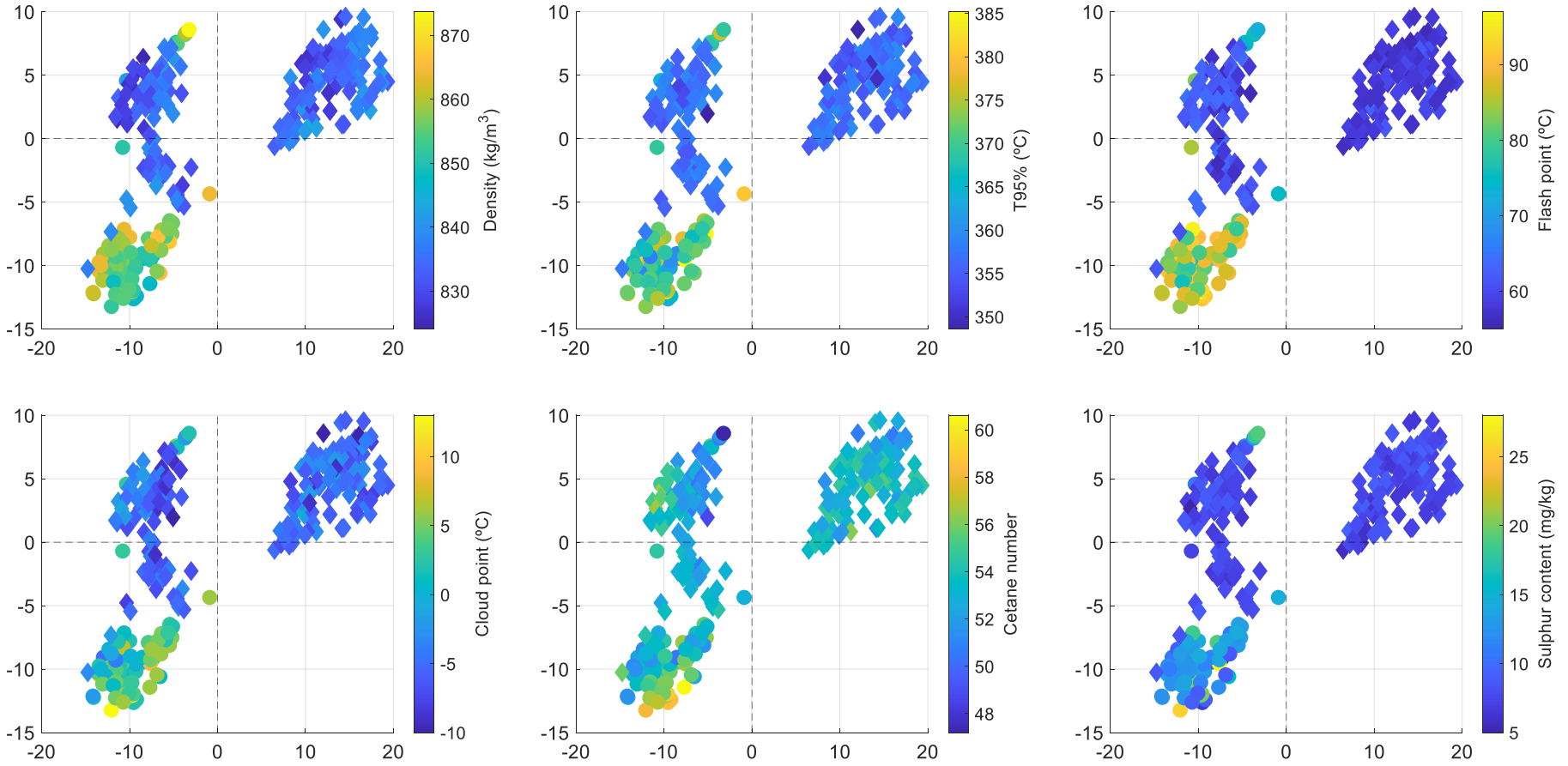


Figure III-9. Plot of the spectral data {stream 1 (circles) and stream 2 (rhombus)} after t-SNE by using cosine distance metrics, perplexity = 30, learning rate = 500, and the reference values of the common properties for both streams in each sample (colour bar).

III.5 Conclusions

For samples from stream 1, no significant correlations between the physicochemical properties were identified. For samples from stream 2, stronger correlations between T65% with T85% (r = 0.90) and T85% with T95% (r = 0.876) were identified.

Samples containing FAME can be identified by the absorption band around 1750 cm^{-1} corresponding to the carbonyl group C=O. There are no specific absorption bands in the infrared region for other properties of diesel.

Principal component analysis of the spectra differentiated the samples into two clusters based on their FAME content. These groups do not correspond to the samples from each stream.

Exploratory analysis using t-SNE proved to be more effective than PCA in distinguishing clusters according to their spectral features. Through this approach, samples from stream 1, samples from stream 2 with FAME content, and samples from stream 2 without FAME content were identified.

The difference in the property values of the samples seems to be more influenced by the origin of the samples (streams 1 or 2) rather than by the FAME content. This observation has led to the development of specific calibration models for samples from each stream.

III.6 References

[1] T.C. Zannis, D.T. Hountalas, R.G. Papagiannakis, Y.A. Levendis, Effect of fuel chemical structure and properties on diesel engine performance and pollutant emissions: Review of the results of four european research programs, SAE Int. J. Fuels Lubr. 1 (2009) 384–419. https://doi.org/10.4271/2008-01-0838.

[2] J. Bacha, J. Freel, A. Gibbs, L. Gibbs, G. Hemighaus, K. Hoekman, J. Horn, M. Ingham, L. Jossens, D. Kohler, D. Lesnini, J. McGeehan, M. Nikanjam, E. Olsen, R. Organ, B. Scott, M. Sztenderowicz, A. Tiedemann, C. Walker, J. Lind, J. Jones, D. Scott, J. Mills, Diesel Fuels Technical Review, 2007. https://www.chevron.com/-/media/chevron/operations/documents/diesel-fuel-tech-review.pdf.

[3] V. Bazán Salazar, Calidad del diesel de Costa Rica entre los años 2006-2010, Cienc. y Tecnol. 31 (2016) 37–51.

[4] M.M. del C. Fernandez-Feal, L.R. Sánchez-Fernández, B. Sánchez-Fernández, Distillation: Basic Test in Quality Control of Automotive Fuels, in: M.F. Mendes (Ed.), Distill. - Innov. Appl. Model., InTechOpen, Rijeka, 2012: pp. 77–97. https://doi.org/10.5772/67140.

[5] P. Saxena, S. Jawale, M.H. Joshipura, A review on prediction of properties of biodiesel and blends of biodiesel, Procedia Eng. 51 (2013) 395–402. https://doi.org/10.1016/j.proeng.2013.01.055.

[6] R.M.E. Dias, R.T. Aquino, M.A. Krähenbühl, M.C. Costa, Flash Point of Fatty Acid Methyl Ester Binary Mixtures, J. Chem. Eng. Data. 64 (2019) 3465–3472. https://doi.org/10.1021/acs.jced.9b00267.

[7] M.F. Ferrão, M.D.S. Viera, R.E.P. Pazos, D. Fachini, A.E. Gerbase, L. Marder, Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions, Fuel. 90 (2011) 701–706. https://doi.org/10.1016/j.fuel.2010.09.016.

[8] S.M. Santos, D.C. Nascimento, M.C. Costa, A.M.B. Neto, L. V. Fregolente, Flash point prediction: Reviewing empirical models for hydrocarbons, petroleum fraction, biodiesel, and blends, Fuel. 263 (2020) 116375. https://doi.org/10.1016/j.fuel.2019.116375.

[9] J.C.L. Alves, R.J. Poppi, Simultaneous determination of hydrocarbon renewable diesel, biodiesel and petroleum diesel contents in diesel fuel blends using near infrared (NIR) spectroscopy and chemometrics, Analyst. 138 (2013) 6477–6487. https://doi.org/10.1039/c3an00883e.

[10] J. Workman Jr., L. Weyer, Practical Guide and Spectral Atlas for Interpretive Near-Infrared Spectroscopy, 2nd ed., CRC Press Taylor & Francis Group, Boca Raton, Florida, 2012.

[11] L. de F. Bezerra de Lira, F.V. Cruz de Vasconcelos, C. Fernandes Pereira, A.P. Silveira Paim, L. Stragevitch, M.F. Pimentel, Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration, Fuel. 89 (2010) 405–409. https://doi.org/10.1016/j.fuel.2009.05.028.

[12] M.G. Nespeca, R.R. Hatanaka, D.L. Flumignan, E. De Oliveira, Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis, J. Anal. Methods Chem. (2018) 1795624. https://doi.org/10.1155/2018/1795624.

[13] N.B. Colthup, L.H. Daly, S.E. Wiberley, Introduction to Infrared and Raman

Spectroscopy, Academic Press, Cambridge, MA, USA, 1990. https://books.google.es/books?id=SDNRAAAAMAAJ.

[14] M.C. Breitkreitz, I.M. Raimundo, J.J.R. Rohwedder, C. Pasquini, H.A. Dantas Filho, G.E. José, M.C.U. Araújo, Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration, Analyst. 128 (2003) 1204–1207. https://doi.org/10.1039/b305265f.

[15] R.M. Silverstein, F.X. Webster, D. Kiemle, Spectrometric Identification of Organic Compounds, 7th ed., Jhon Wiley & Sons, Inc, United States of America, 2005. https://books.google.es/books?id=mQ8cAAAAQBAJ.

[16] M. Fernanda Pimentel, G.M.G.S. Ribeiro, R.S. Da Cruz, L. Stragevitch, J.G.A. Pacheco Filho, L.S.G. Teixeira, Determination of biodiesel content when blended with mineral diesel fuel using infrared spectroscopy and multivariate calibration, Microchem. J. 82 (2006) 201–206. https://doi.org/10.1016/j.microc.2006.01.019.

[17] B. Melit Devassy, S. George, Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE, Forensic Sci. Int. 311 (2020) 110194. https://doi.org/10.1016/j.forsciint.2020.110194.

[18] B.M. Devassy, S. George, P. Nussbaum, Unsupervised clustering of hyperspectral paper data using T-SNE, J. Imaging. 6 (2020) 29. https://doi.org/10.3390/JIMAGING6050029.

Chapter IV
Calibration model based on artificial neural network for density prediction. Defining the limits of its applicability domain[1]

[1] Adapted from M. S. Rodríguez Barrios, J. Ferré, M.S. Larrechi, E. Ruiz, *Applicability domain of a calibration model based on neural networks and infrared* spectroscopy, Chemom. Intell. Lab. Syst. (Submitted).

Chapter IV

IV.1 Introduction

The demand for fast, inexpensive, and multiparametric analytical methods has led to the widespread use of infrared spectroscopy with multivariate calibration. Among the calibration models, artificial neural networks (ANNs) are a versatile class that is especially suited to handle nonlinear relationships between the spectrum and the property to be predicted [1]. Some examples of analytical determinations that use ANNs can be found in references [1-7].

The use of ANNs in routine analyses requires proper model validation, the specification of the applicability domain, the periodic testing of the model's performance against the reference method, and some means of estimating the prediction uncertainty. While aspects such as prediction uncertainty [8-9] and analytical figures of merit [9-11] have been addressed, the characterization of the applicability domain of a feed-forward neural network has been less studied. The applicability domain (AD) can be defined as the sample space where the model is expected to predict with a given reliability [12]. In spectroscopic methods, predictions are more reliable when the spectra of the new samples are like the spectra of the training samples, i.e., within the AD. Samples outside the AD are extrapolations, and therefore, their predictions are less reliable. The reasons why a new sample may fall outside the AD are diverse. Some are related to the sample itself, such as having a new ratio of constituents or different matrix effects, while others are related to gross errors, new measurement conditions, or instrument performance. These sources of discrepancy are discussed regularly in reports that deal with novelty detection, anomaly detection, fault detection, and model updating. In this sense, the specification of the AD is connected to the approaches used to detect prediction outliers, which are those instances that are outside the boundaries of the AD.

Quantitative-structure activity relationship (QSAR) studies offer a convenient initial perspective about some approaches that can be used to describe the AD. The definition of the AD in QSAR is an active area of research [13-18] motivated by the fact that the validation of QSAR models used in regulatory applications must include the specification of the AD [19]. When the modeling algorithms used in analytical determinations coincide with those in QSAR, the principles used to define the AD are common [20-21]. These principles are also found in surveys about outlier detection in classification and regression [22-28], where the AD is not explicitly mentioned, but it is indirectly understood as the space enclosed by the limits used to issue outlier warnings.

There are different approaches to defining the AD of a model. They can be based on variable ranges, distances, densities, ensemble learning, or prediction intervals, to

mention one way of grouping them [20,29]. This diversity of methods reflects the variety of scenarios, and the choice of the approach depends on the characteristics of the model and the type and distribution of the available data. Nevertheless, the prevailing situation is the lack of training instances outside the AD, so most methods are unsupervised, and the boundaries of the AD are set by learning from good data. Among these methods, distance-based approaches are simple and widely used [13,17,18,30]. The similarity between a new sample and the training samples can be measured by the Euclidean distance [14], Mahalanobis distance [31], Manhattan distance [32], Hotelling's T^2 and leverage [29]. For highly correlated variables, such as spectra, the Mahalanobis distance is especially useful [33,34]. It is assumed that the prediction accuracy decreases as the distance of the new sample to the training samples increases. Hence, a distance above a given threshold indicates that the sample is outside the AD, and its prediction is unreliable. The threshold can be set as a quantile of the theoretical distribution that the metric is known to follow or a quantile of the distances calculated for the training samples up to the maximum distance obtained for the training set [13,29,35]. Reconstruction-based methods are also common. They assume that novelties (i.e., spectra of good samples not yet represented by the model) or anomalies (i.e., spectra of bad samples or erroneous spectra of good samples) cannot be reconstructed from projections in a space of fewer dimensions calculated from good samples [25,36]. Therefore, they can be detected by a spectrum reconstruction error above a given threshold [21]. This idea is used in factor-based multivariate models, where the sum of squared spectral residuals (a.k.a. Q-value) is commonly used to detect outliers. In nonlinear scenarios, autoencoders are a type of neural network that can be used for this purpose. These networks are designed to reconstruct the input through a narrow hidden layer known as the bottleneck. By compressing the input through the narrow layer, the common information relevant to describing the samples is preserved, so the autoencoder learns to reconstruct the training data approximately. Hence, an autoencoder will properly reconstruct the spectrum of a new sample if it is similar to the spectra used to train it. Otherwise, reconstruction will be poor. Therefore, novel and anomalous samples can be identified by having larger spectral residuals than the residuals of the training samples. Autoencoders have been used in multivariate process control, sample classification, and fault diagnosis, among other applications [7,37].

This work shows a procedure for establishing the AD when a feed-forward neural network (FFNN) is the calibration model of an analytical determination. The AD is enclosed by two limits. One limit is based on the Mahalanobis distance calculated from the projection of the training spectra in the reduced latent space of the regression

network. The other limit uses the sum of squared spectral residuals of an autoencoder. Two variations of the latter are presented and compared. One is the classical autoencoder, trained from the raw spectra. The other uses only the decoder part of an autoencoder, trained using the activations from the regression network as input.

As a study case, the approach is applied to the determination of the density of diesel fuel samples from IR spectra using an FFNN. The combination of IR spectroscopy and ANNs has been shown to be effective in determining relevant physicochemical properties of diesel fuel samples [38-41].

IV.2 Theory

IV.2.1 Autoencoder in anomalies detection

Autoencoders can be used to detect anomalies since, by forcing the output to be as similar as possible to the input, they generalize poorly. An autoencoder trained with the spectra that were used for training the regression network will faithfully reconstruct a new sample spectrum only if it resembles the training spectra. An anomalous spectrum containing unmodeled parts will not be reproduced well, and the spectral residuals will be large. A threshold $Q_{\lim}$ can be estimated from the spectral residuals of the training spectra so that a new spectrum whose Q value is larger than $Q_{\lim}$ is said to be outside the applicability domain of the model. The threshold will depend on the complexity of the autoencoder and the similarity among the training spectra. For a given autoencoder architecture, a training set that is made of very similar spectra will result in a well-fitted autoencoder, small residuals and a low $Q_{\lim}$. Only very similar new spectra will be reproduced well and the applicability domain will be restricted. A slightly different spectrum will easily be marked as outside the applicability domain. This implies an increased risk of rejecting a good extreme sample (type I error). Conversely, an autoencoder trained with very diverse spectra will have a worse fit and a higher $Q_{\lim}$. The applicability domain will be less tight and more varied spectra will be accepted, increasing the likelihood that a true outlier will go unnoticed (type II error). The number of layers and nodes of the autoencoder will also affect $Q_{\lim}$. For a given training set, a wider bottleneck results in a better fit and a lower $Q_{\lim}$ than a narrow bottleneck. Thus, fine-tuning the autoencoder architecture can restrict or widen the applicability domain. The trade-off will depend on the problem at hand.

Note that an autoencoder will reproduce a spectrum independently on the regression network for which we are specifying the applicability domain. In this sense, defining the applicability domain of a regression network using the autoencoder's ability to reproduce spectra is like using the nearest neighbor distance or principal component analysis to

define the applicability domain of a partial least squares regression model. None of these methods use the specific form of the regression model to indicate that a spectrum is inside or outside the applicability domain. Since the regression network emphasizes different parts of the spectrum depending on the property to be predicted, the effect of a spectral anomaly on the prediction is different depending on the zone of the spectrum affected. To emphasize the detection of anomalies in the wavelength ranges that are most influential for prediction, an alternative spectral reconstruction approach can be tested. It consists of training only the right-hand block of the autoencoder, the decoder, to reproduce the training spectra. In this case, the input to the decoder for the training sample i can be the activations vector of the hidden layer of the regression network $\mathbf{a}_i$ and the output will be the reconstructed spectrum $\hat{\mathbf{x}}_i$. The decoder is then trained to minimize the reconstruction error given by equation II.30 in Chapter II. This is a less optimal implementation of an autoencoder, as only the decoder part is trained starting from a loose representation of the spectrum. It results in larger spectral residuals but is an attempt to increase the sensitivity to anomalies at the wavelengths that are relevant for regression.

IV.2.2 Mahalanobis distance

Let $\boldsymbol{X} = [X_1, \ldots, X_p]^{\mathrm{T}}$ be $p \times 1$ random vector with population mean $\boldsymbol{\mu} = E(\boldsymbol{X})$ and covariance matrix $\boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathrm{T}}]$. The Mahalanobis distance between $\boldsymbol{X}$ and $\boldsymbol{\mu}$ is given by

$$D(\boldsymbol{X}, \boldsymbol{\mu}) = \{(\boldsymbol{X} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})\}^{1/2} \quad \text{(IV.1)}$$

This scalar is a generalized distance that measures where a vector $\boldsymbol{X}$ lies with respect to the center of the multivariate space taking into account the correlations among variables. If $\boldsymbol{X}$ is normally distributed, i.e., $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, D^2 follows a chi-square distribution with p degrees of freedom [42]. In practice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the M training samples as $\bar{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_i$ and $\mathbf{S} = \frac{1}{M-1}\sum_{i=1}^{M} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$ so the sample version of the squared Mahalanobis distance for observation $\mathbf{x}_i$

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}), \qquad i = 1, \ldots, M \quad \text{(IV.2)}$$

is only approximately distributed as a χ_p^2 if $\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ but will asymptotically approach χ_p^2 as M gets larger.

In linear calibration models such as multiple linear regression, principal components regression or partial least squares regression, the Mahalanobis distance is related to the leverage, which is a recommended measure to flag that a new sample spectrum is outside the calibration range [43]. The Mahalanobis distance can be used for the same purpose as the leverage, by setting a limit of the AD as a quantile (e.g., 0.99) of χ_p^2 in case of multivariate normality or a quantile of the D^2 values of the training set if $\boldsymbol{X}$ is not normally distributed. This latter approach resembles the ASTM norm recommendation that a leverage larger than the maximum leverage of the calibration set can be used as an indication of extrapolation of the model.

Although the Mahalanobis distance can be calculated for the spectra, it requires the covariance matrix be nonsingular, which limits its use to cases when the number of samples is larger than the number of variables. Moreover, the fitted model is not considered to define its AD because eq IV.2 only involves the spectra. A less restrictive option is to calculate the Mahalanobis distance using the activations of the hidden layer of the regression network as

$$D_i^2 = (\mathbf{a}_i - \overline{\mathbf{a}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{a}_i - \overline{\mathbf{a}}) \qquad \text{(IV.3)}$$

where $\mathbf{a}_i$ is the vector of activations of the hidden layer for sample i and $\overline{\mathbf{a}}$ and $\mathbf{S}$ are the average and covariance matrix of the hidden layer activations of the training set respectively. As noted above, if the underlying distribution of the activations is multinormal, then a quantile of χ_d^2 where d is equal to the dimension of $\mathbf{a}$ (that is, the number of nodes in the hidden layer) can be used as a limit of the applicability domain. If D^2 does not follow a chi-square distribution, then a limit D_{lim}^2 can be set as a quantile (e.g., 0.99) of the D^2 values of the training set. In all cases, a higher D_i^2 than the limit will indicate that a sample outside the applicability domain and that its prediction cannot be trusted [44]. This work describes the case when the regression network has only one hidden layer. Such a model was enough for the property being modelled. For regression networks with more than one hidden layer, one could hypothesize that combining the Mahalanobis distances calculated for each hidden layer could be another way of defining the applicability domain. This hypothesis has not been considered in this work.

IV.2.3 Q spectral residual

The root mean square of spectral residuals has been recommended [43] to detect that a new sample contains interferences not present in the calibration samples.

Similarly, for an autoencoder that has been trained to reproduce an input spectrum, define the Q value for a sample [45] as the squared reconstruction error of the spectrum:

$$Q = \sum_{k=1}^{K} (x_k - \hat{x}_k)^2 \qquad \text{(IV.4)}$$

A new sample will be considered to be outside the AD if the sum of squared spectral residuals from the autoencoder is larger than a limit set from the spectral residuals of the training data. If the spectra follow a multinormal distribution, then Q follows a χ^2_K distribution with the number of degrees of freedom equal to the number of spectral variables K. In that case, a limit Q_{lim} can be set as a certain quantile (e.g., 0.99) of the chi-square distribution. Otherwise, Q_{lim} can be set, for example, as the 0.99 quantile of the Q values of the training set. Thus, a new sample is said to be outside the applicability domain if $Q > Q_{\text{lim}}$.

IV.2.4 Applicability domain of a regression network

The applicability domain of the regression neural network can be established as follows. Once a regression network has been trained to predict a property of interest from the infrared spectrum, the vector of activations of each training sample in the hidden layer of the regression network $\mathbf{a}_i$ (Figure II-12) is kept aside. The activations of all the training samples are then used to calculate $\mathbf{S}$ in equation IV.3, the squared Mahalanobis distance of each training sample, and the limit of the applicability domain D^2_{lim}. Next, two approaches are proposed to calculate the limit of the applicability domain that is based on the spectral residuals, Q_{lim}. The first approach consists of training an autoencoder (Figure II-13) to reconstruct the training spectra using the training spectra as input. The spectral residuals of the training set are used to define Q_{lim}. Alternatively, only the decoder part of Figure II-13 is trained to reproduce the training spectra using as input the activations $\mathbf{a}_i$ that were used to calculate the Mahalanobis distance. The spectral residuals are used to define Q_{lim}. The spectrum of the new sample is submitted to the regression network to obtain the prediction and the Mahalanobis distance is calculated from the activation of the hidden layer. Next, either the spectrum is submitted to the autoencoder to obtain the reconstructed spectrum $\hat{\mathbf{x}}$ and Q (equation IV.4) or the hidden layer activations are submitted to the decoder to obtain the reconstructed spectrum $\hat{\mathbf{x}}$ and Q. The spectrum is inside the applicability domain if $Q \leq Q_{\text{lim}}$ and $D^2 \leq D^2_{\text{lim}}$.

Otherwise, the prediction is not reliable and must be verified using the reference analytical method.

IV.3 Matherials and methods

IV.3.1 Samples and software

The sample set consisted of 1792 diesel samples from streams 1 and 2 described in section III.1 of Chapter III. Absorbance spectra were acquired using the procedure described in section III.4.1 of Chapter III. Absorbance spectra between 2591.86-1785.76 cm^{-1} were considered for the development of the global calibration model. The density of the samples was measured following the ASTM D4052 method [46] with an ANTON PAAR digital densimeter model DMA 4500M. The values ranged between 820.0 kg/m^3 and 875.0 kg/m^3. The dataset was randomly split into a training set (899 samples), a validation set (451 samples) and a test set (434 samples) which contained 50%, 25% and 25% of the samples analyzed each month over 35 months.

Five spectra measured in a second FTIR/FT-NIR instrument of the same manufacturer were added to the test set. Three discordant spectra resulting from an erroneous manipulation of the sample in the instrument and a flawed background measurement were also added to the test set.

Calculations were carried out in MATLAB (The MathWorks Inc., Natick, MA, R2022a) with MATLAB's Deep Learning Toolbox™ and homemade routines.

IV.3.2 Methodology for optimizing artificial neural networks

A regression network was trained to predict the density of the diesel samples from the mid-infrared spectra. The type of neural network used to predict diesel quality parameters was a FFNN. The network consisted of a combination of layers as described in the Deep Learning toolbox of Matlab software: an input layer that received the spectral data after, min-max normalization (see Chapter II) preprocessing, a hidden layer with the hyperbolic tangent activation function, and an output layer with a linear activation function (see Figure II-12 in Chapter II). The architecture of the network was selected after training different networks with one and two hidden layers, with different combinations of number of nodes in each layer up to 25 nodes. The models were trained by backpropagation with 5000 or more epochs to minimize the MSE with L_2-regularization. Training used stochastic gradient descent with momentum, initial learning rate 0.001,

and L_2-regularization factor 10^{-3}. A mini-batch size of 16 was used to evaluate the gradient of the loss function and update the weights. The simplest network with a low MSE in the validation set was selected.

Different setups of autoencoder with three hidden layers were tested with an increasing number (up to 25) of neurons each. To reduce the number of possible structures that could be evaluated, the number of neurons in the bottleneck layer was the number of nodes in the hidden layer of the selected regression network. The rest of the settings for training the autoencoder were the same as those used for the regression network. For the decoder, different setups of decoder were tested. Except L_2-regularization factor 10^{-3}, the training parameters for the decoder were the same as those used for the autoencoder. Both in the autoencoder and the decoder, the number of neurons of the output layer was the dimensions of the spectrum.

IV.4 Results and discussion

IV.4.1 Neural network models

Figure IV-1 shows the MIR region used to establish the regression model. The range 2591.86-1785.76 cm^{-1} contains the fundamental vibrations of groups C-H and C=O present in most compounds in diesel samples [47]. Regression networks with one or two hidden layers with a variable number of nodes were trained to predict density from MIR spectra. The selected model was the network with one hidden layer (Figure II-12) and ten neurons, as it was the simplest model with an acceptable prediction error of the validation set. Only small differences in the validation performance were found by using more than ten neurons or a second hidden layer. Such improvements fluctuated depending on the randomness of the iterative training process based on minibatches. Figure IV-2 shows the predicted density for the calibration and validation sets using the selected network. The root mean squared error (RMSE) for the calibration and validation sets was 0.56 kg/m^3 and 0.62 kg/m^3 respectively. The determination coefficient (R^2) of the linear fit between the reference density and the predictions was 0.98 and 0.98 for the training and validation sets, respectively. The prediction ability of the network was comparable to previously reported results [39] with ANNs or other multivariate calibration methods calculated with smaller sets of samples covering shorter production periods.

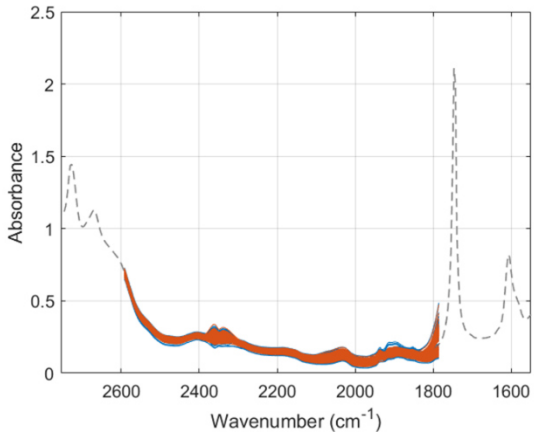


Figure IV-1. MIR spectra of the diesel samples: training set (blue) and validation set (orange).

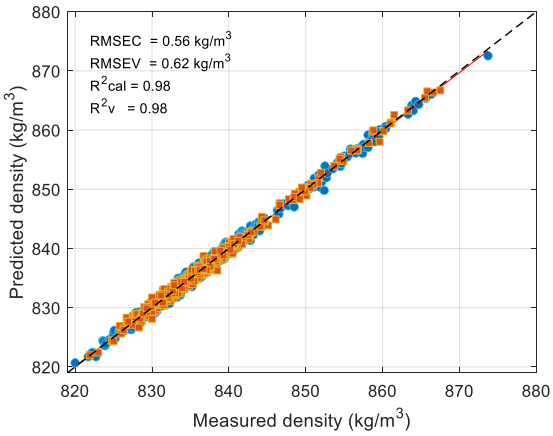


Figure IV-2. Predicted versus reference density for the training (blue) and validation (orange) samples.

Different autoencoder architectures with one or three hidden layers were tested to reproduce the training spectra. To keep the number of combinations to be tested low, the number of neurons in the bottleneck was fixed, and it was the same as the number of neurons in the hidden layer in the regression network. The selected autoencoder had 15, 10, and 15 neurons in the first, middle (bottleneck), and third hidden layers, respectively. The RMSE for the training and validation sets in the selected model was $8.21 \cdot 10^{-4}$ and $8.47 \cdot 10^{-4}$, respectively. Figure IV-3 shows the reconstructed spectra $\hat{\mathbf{x}}$ and the spectral residuals of the calibration and validation samples. For both the calibration and validation sets, the correlation coefficient between each spectrum and the reconstructed spectrum ranged from 0.9992 to 1, confirming that the autoencoder could successfully reproduce the spectra.

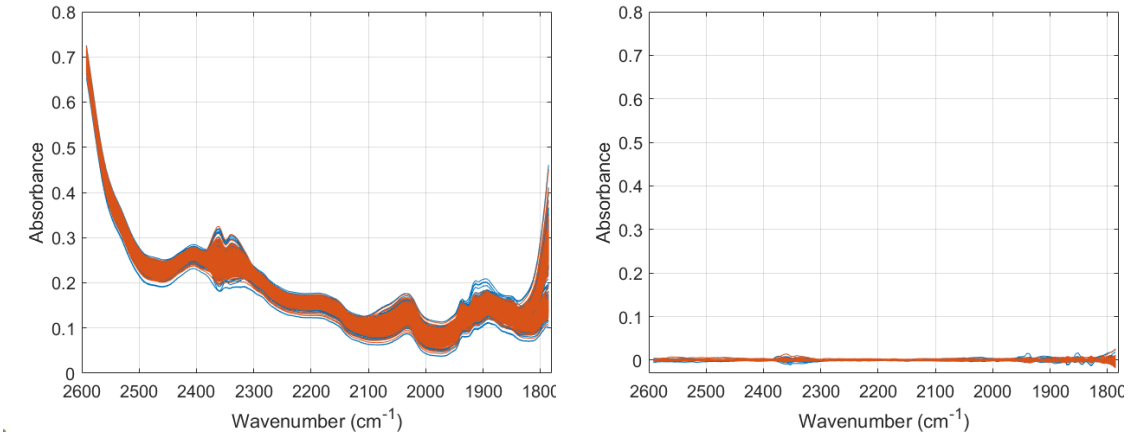


Figure IV-3. Reconstructed training spectra and spectral residuals from the autoencoder.

The decoder was trained using as inputs the hidden layer activations of the training set from the regression network (a 10×1 data vector per sample). Like in the autoencoder, the output layer returned the reconstructed spectrum and equation II.30 in Chapter II with L_2-regularization was the loss function to be minimized. Fifteen different architectures of the decoder were tested with one and two hidden layers, with an increasing number of neurons, up to 25. The simplest network with one hidden layer and 15 neurons was selected as it provided the lowest RMSE for the validation set. The reconstruction error did not improve by using two hidden layers. The RMSE for the training and validation sets in the selected model was $1.43 \cdot 10^{-3}$ and $1.42 \cdot 10^{-3}$, respectively. Figure IV-4 shows the spectra estimated by the decoder for the training and validation sets and the spectral residuals. For both the calibration and validation sets, the correlation coefficient between each spectrum and the estimated spectrum ranged between 0.9986 and 1, confirming that the decoder could also correctly reproduce the original spectra from the ten activation values obtained from the regression network. The spectral residuals are larger than the residuals produced by the autoencoder, which is consistent with the fact that the autoencoder has a more complex structure with more coefficients and is thus expected to fit the training spectra better.

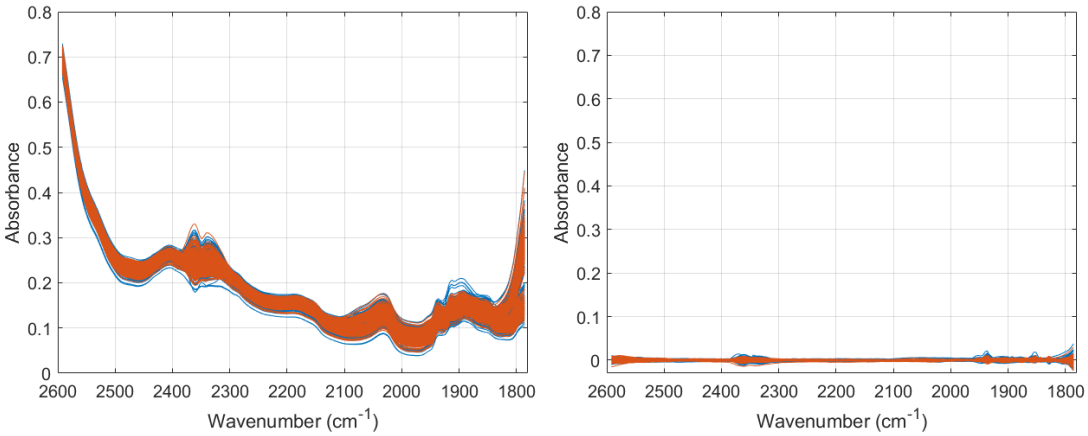


Figure IV-4. Reconstructed training spectra and spectral residuals from the decoder.

IV.4.2 Applicability domain of the regression network

The squared Mahalanobis distance of calibration, validation and test spectra was calculated from the activations of the hidden layer of the regression network (equation IV.3). The quantile-quantile plot [48] of the D^2 values of the training samples (Figure IV-5) indicated that D^2 does not follow a χ^2_{10} distribution, so the limit of the applicability domain D^2_{lim} was set at the 0.99 quantile of the D^2 values of the training set, which was 39.49.

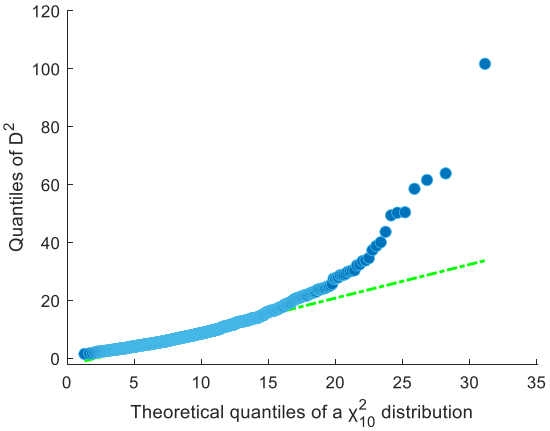


Figure IV-5. Quantiles of D^2 vs. theoretical quantiles of a χ^2_{10} distribution.

The Q values were calculated from the spectral residuals of the autoencoder and the decoder. The quantile-quantile plot indicated that Q for both the autoencoder (Figure IV-6) and the decoder (Figure IV-1S in the supplementary information) do not follow a chi-square distribution. For the autoencoder, the largest Q value was $3.60 \cdot 10^{-3}$ and the limit $Q_{\text{lim}-\text{AE}}$ was set at $1.20 \cdot 10^{-3}$, the 0.99 quantile of the Q values of the training set. For the decoder, the largest Q value of the training set was $6.60 \cdot 10^{-3}$ and $Q_{\text{lim}-\text{EN}}$ was

set at $2.70 \cdot 10^{-3}$, the 0.99 quantile of the Q values of the training set. Note that $Q_{\mathrm{lim-AE}}$ is lower than $Q_{\mathrm{lim-EN}}$, which is consistent with the smaller spectral residuals of the autoencoder observed above.

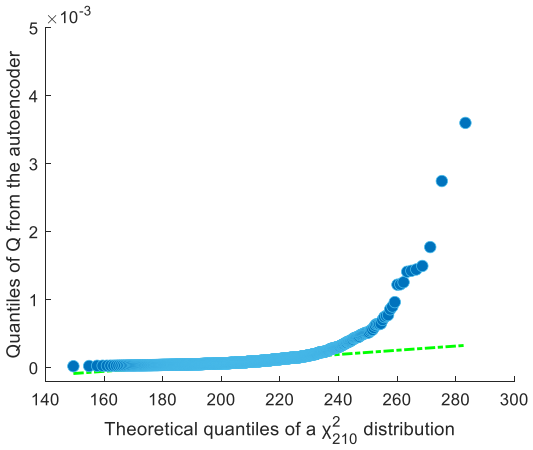


Figure IV-6. Quantiles of Q from the autoencoder vs. theoretical quantiles of a χ^2_{101} distribution.

Figure IV-7 shows the limits of the AD of the regression network defined by the squared Mahalanobis distance and the spectral residuals of the autoencoder. 1% of the training samples (that is, 9 samples) had D^2 values higher than D^2_{lim} or Q values higher than $Q_{\mathrm{lim-EN}}$. In total, 15 training samples (1.7% of the set) are outside the applicability domain because they exceed one of the two limits D^2_{lim} or $Q_{\mathrm{lim-AE}}$ or even both. This is an expected consequence of selecting 0.99 quantiles to set the limits of the AD. The presence of these valid samples outside the AD indicates that the limits of the applicability domain should not be interpreted as rigid boundaries that separate good samples from outliers with unreliable predictions. Rather, these limits are the borders of an inner region where the predictions can be confidently accepted.

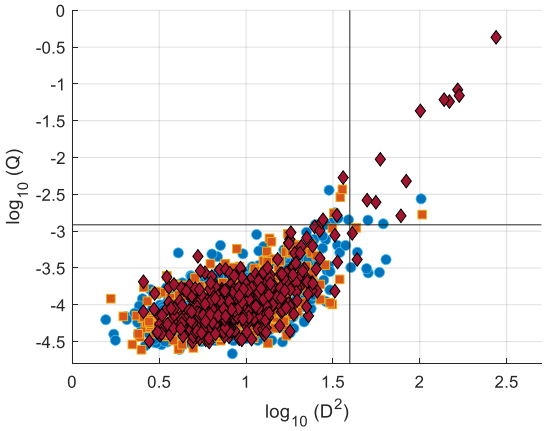


Figure IV-7. $log_{10}(Q)$ of the autoencoder vs. $log_{10}(D^2)$ for the training (blue), validation (orange) and test samples (red). The limits of the applicability domain of the regression network are shown. The logarithmic scale was used to facilitate the visualization.

Only four validation samples had Q values higher than $Q_{\mathrm{lim-AE}}$ and one of them also had D^2 value higher than D^2_{lim}. Most of the test samples were also within the AD, except 19 samples that had D^2 or Q values exceeding the limits. Four of them stood out only for their high spectral residual ($Q > Q_{\mathrm{lim-AE}}$) while their D^2 values were lower than D^2_{lim}. Two samples stood out for their high D^2 values ($D^2 > D^2_{\mathrm{lim}}$) while having a low spectral residual ($Q < Q_{\mathrm{lim-AE}}$). The thirteen remaining samples had simultaneously D^2 and Q values exceeding the limits. Figure IV-8 shows the spectra of these thirteen highly discordant samples, compared with the spectra of the training samples. Among these thirteen, five were the ones that had been measured with the second instrument and three were the rare spectra that had been added to the test set. This showed the ability of the AD to flag discordant spectra.

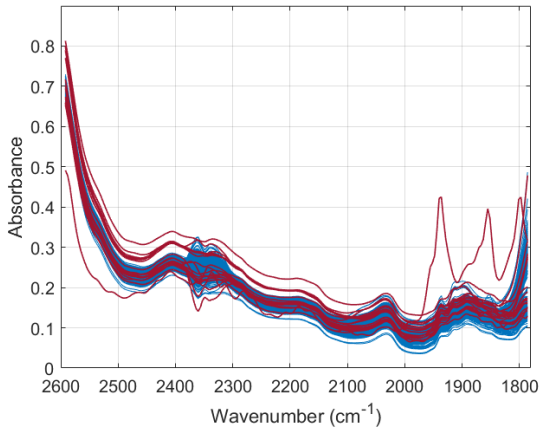


Figure IV-8. Spectra of the samples outside the applicability domain compared to the range of the spectra of the training samples.

Figure IV-9 shows the empirical cumulative distribution function of the absolute prediction error of the density for each sample $e_i = |y_i - \hat{y}_i|$ of the training and validation set together. As expected for regular spectra, there is not a high correlation between the absolute prediction error and the $D^2$and Q values as they all are part of the modelled variability of the training set. Hence, a range of prediction errors is possible, as indicated by the cumulative curve. On the other hand, anomalous spectra that have high values of Q and D^2 (that is, they are outside the AD) are more susceptible to produce large prediction errors. This is what is observed in Figure IV-9 with the ten test samples that were outside both AD limits simultaneously. Nine of them had absolute errors larger than $1.35\ \mathrm{kg/m^3}$, that is, worse than 97% of the errors of the training and validation sets. The other four test samples that were outside the AD had small absolute errors, in accordance with the idea that the AD cannot be regarded as the space outside which we have the absolute certainty that the prediction errors will be large in all the cases, but as

a zone outside which there is a higher risk of large prediction errors and thus, the predictions should not be trusted.

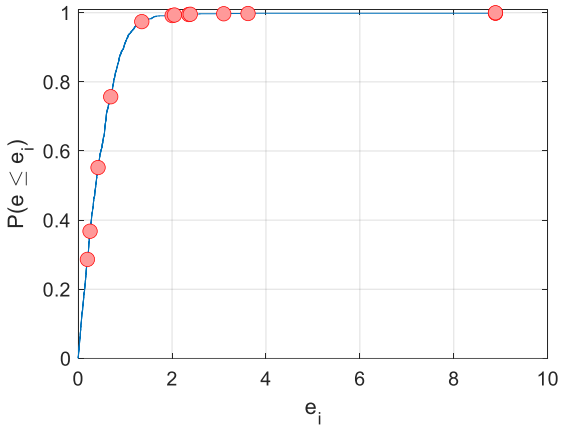


Figure IV-9. Empirical cumulative distribution function of the absolute prediction error for density.

As an alternative to the autoencoder, which is trained independently on the regression network, the spectral residual limit of the AD was also calculated from the decoder part of an autoencoder, calculated from the activations of the hidden layer of the regression network. The decoder reconstructs the spectra worse than the autoencoder, resulting in larger residuals and a larger Q_{lim}. The AD limits set from the Mahalanobis distance and the decoder also detected the thirteen discordant spectra in the test set with a Q value larger than Q_{lim} that were also flagged by the autoencoder (Figure IV-2S in the supplementary information), but we did not observe any apparent improvement in the detection of outlying samples by using the decoder compared to using the autoencoder. The reason was that the decoder did not necessarily had the largest weights at the nodes that received the largest (in absolute value) activations from the hidden layer of the regression network. Therefore, small anomalies at the important wavelengths that resulted in defective activations (and hence, could be detected by the Mahalanobis distance) did not translate into large spectral residuals in a more sensitive way than for the autoencoder.

IV.5 Conclusions

Routine quantitative determinations based on multivariate spectroscopy and multivariate calibration require safeguards to ensure that the accepted size of the prediction errors is maintained during the use of the model. One of these safeguards is the characterization of the applicability domain of the model. Beyond the limits of this domain, the prediction of a sample is not considered to be reliable enough. This does not necessarily mean that the sample will produce a large prediction error, but it is susceptible to it. This work has presented an approach to finding these limits in a feed-forward neural network, arguably the most common and simple network used in chemical analysis. The limits were set as the 0.99 quantiles of two metrics calculated from the training data. One was the squared Mahalanobis distance calculated from the activations of the hidden layer of the regression network. This measured the distance of the new sample to the center of the model. The second was the sum of squared spectral residuals obtained by reconstructing the spectrum from a lower dimensional space. For the latter, two approaches were studied, either using an autoencoder or a decoder. The autoencoder was trained independently on the regression network, while the decoder used the activations from the hidden layer of the regression network. Both approaches indicated how well the prediction spectrum could be reproduced from the training data. Large spectral residuals indicated a spectrum outside the applicability domain and a questionable prediction. As a case study, a neural network was used to predict the density of diesel fuel samples from their MIR spectra. Thirteen spectra were identified outside the applicability domain with Q and D^2 values higher than the limits simultaneously. Nine of them were found to have high prediction errors. For the remaining four samples, although their prediction should not be trusted, the prediction errors were not abnormally high. While the initial hypothesis that the decoder might outperform the autoencoder could not be confirmed, both methodologies showed consistent behavior.

This work has been restricted to a regression network with one hidden layer, but there are a variety of architectures of networks that could be considered. For example, the activations of a regression network with two hidden layers could be combined to offer a better limit of the applicability domain. The idea that a better definition of the applicability domain should involve the form of the model (something that is not taken into account by the autoencoder only) still persists.

IV.6 References

[1] F. Despagne, D. L. Massart, Neural networks in multivariate calibration. Analyst 123 (1998) 157R–178R. doi:10.1039/A805562I.

[2] D.A. Cirovic, Feed-forward artificial neural networks: applications to spectroscopy, Trends Anal. Chem. 16 (1997) 148–155. doi:10.1016/S0165-9936(97)00007-1.

[3] D. Pérez-Marín, A. Garrido-Varo, J.E. Guerrero, Non-linear regression methods in NIRS quantitative analysis, Talanta 72 (2007) 28–42. doi:10.1016/j.talanta.2006.10.036.

[4] F. Marini, Artificial neural networks in foodstuff analyses: Trends and perspectives. A review, Anal. Chim. Acta 635 (2009) 121–131. doi:10.1016/j.aca.2009.01.009.

[5] J. Yang, J. Xu, X. Zhang, C. Wu, T. Lin, Y. Ying, Deep learning for vibrational spectral analysis: Recent progress and a practical guide, Anal. Chim. Acta, 1081 (2019) 6–17. doi:10.1016/j.aca.2019.06.012.

[6] Y. Chen, L. Song, Y. Liu, L. Yang, D. Li, A review of the artificial neural network models for water quality prediction, Appl. Sci. 10 (2020) 5776–5825. doi:10.3390/app10175776.

[7] P. Mishra, D. Passos, F. Marini, J. Xu, J.M. Amigo, A.A. Gowen, J.J. Jansen, A. Biancolillo, J.M. Roger, D.N. Rutledge, A. Nordon, Deep learning for near-infrared spectral data modelling: Hypes and benefits, Trends Anal. Chem. 157 (2022) 116804–116829. doi: 10.1016/j.trac.2022.116804.

[8] L. Yang, T. Kavli, M. Carlin, S. Clausen, P. F. De Groot, An evaluation of confidence bound estimation methods for neural networks. In: Advances in Computational Intelligence and Learning. International Series in Intelligent Technologies. Dordrecht, The Netherlands: Springer, 18 (2002) 71–84. https://doi.org/10.1007/978-94-010-0324-7_5.

[9] F. Allegrini, A. C. Olivieri, Sensitivity, prediction uncertainty, and detection limit for artificial neural network calibrations, Anal. Chem. 88 (2016) 7807−7812. doi:10.1021/acs.analchem.6b01857.

[10] F.A. Chiappini, F. Allegrini, H.C. Goicoechea, A.C. Olivieri, Sensitivity for multivariate calibration based on multilayer perceptron artificial neural networks, Anal. Chem. 92 (2020) 12265–12272. doi:10.1021/acs.analchem.0c01863.

[11] K. Shariat, D. Kirsanov, A. C. Olivieri, H. Parastar, Sensitivity and generalized analytical sensitivity expressions for quantitative analysis using convolutional neural networks, Anal. Chim. Acta 1192 (2022) 338697-338706. doi:10.1016/j.aca.2021.338697.

[12] N. Fjodorova, M. Novič, A. Roncaglioni, E. Benfenati, Evaluating the applicability domain in the case of classification predictive models for carcinogenicity based on the counter propagation artificial neural network. J. Comput. Aided. Mol. Des. 25 (2011)1147–1158. doi:10.1007/s10822-011-9499-9.

[13] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules. 17 (2012) 4791–4810. doi:10.3390/molecules17054791.

[14] N. Minovski, Š. Župerl, V. Drgan, M. Novič, Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study, Anal. Chim. Acta. 759 (2013) 28–42. doi:10.1016/j.aca.2012.11.002.

[15] M. Mathea, W. Klingspohn, K. Baumann, Chemoinformatic classification methods and their applicability domain, Mol. Inform. 35 (2016) 160–180. doi:10.1002/minf.201501019.

[16] P. Žuvela, J. David, M.W. Wong, Interpretation of ANN-based QSAR models for prediction of antioxidant activity of flavonoids, J. Comput. Chem. 39 (2018) 953–963. doi:10.1002/jcc.25168.

[17] R. Liu, H. Wang, K. P. Glover, M. G. Feasel, A. Wallqvist, Dissecting machine-learning prediction of molecular activity: is an applicability domain needed for quantitative structure-activity relationship models based on deep neural networks?, J. Chem. Inf. Model. 59 (2019) 117–126. doi:10.1021/acs.jcim.8b00348.

[18] Y. Tian, S. Zhang, H. Yin, A. Yan, Quantitative structure-activity relationship (QSAR) models and their applicability domain analysis on HIV-1 protease inhibitors by machine learning methods, Chemometr. Intell. Lab. Syst. 196 (2020) 103888–103902. doi:10.1016/j.chemolab.2019.103888.

[19] Organisation for Economic Co-operation and Development. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models.

OECD Series on Testing and Assessment. 69 (2014) 1-154. doi:10.1787/20777876.

[20] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith, C. Yang, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Altern Lab Anim. 33 (2005) 155–173. doi: 10.1177/026119290503300209.

[21] M.A.F. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (2014) 215–249. doi:10.1016/j.sigpro.2013.12.026.

[22] M. Markou, S. Singh, Novelty detection: a review—part 1: statistical approaches, Signal Process. 83 (2003) 2481–2497. doi:10.1016/j.sigpro.2003.07.018.

[23] M. Markou, S. Singh, Novelty detection: a review: part 2: neural network based approaches, Signal Process. 83 (2003) 2499–2521. doi:10.1016/j.sigpro.2003.07.019.

[24] V.J. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2004) 85–126. doi:10.1007/s10462-004-4304-y.

[25] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (2009) 1–58. doi:10.1145/1541880.1541882.

[26] H. Wang, M. J. Bah, M. Hammad, Progress in outlier detection techniques: A survey, IEEE Access. 7 (2019) 107964–108000. doi:10.1109/ACCESS.2019.2932769.

[27] A. Boukerche, L. Zheng, O. Alfandi, Outlier detection: methods, models, and classification, ACM Comput. Surv. 55 (2020) 1-37. doi:10.1145/3381028.

[28] G. Pang, C. Shen, L. Cao, A. Van Den Hengel, Deep Learning for Anomaly Detection: A Review, ACM Comput. Surv. 54 (2021) 1–36. doi:10.1145/3439950.

[29] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, QSAR applicabilty domain estimation by projection of the training set descriptor space: a review, Altern Lab Anim. 33 (2005) 445-459. doi: 10.1177/026119290503300508.

[30] N. Nikolova, J. Jaworska, Approaches to Measure Chemical Similarity - A Review, QSAR Comb. Sci. 22 (2004) 1006–1026. doi:10.1002/qsar.200330831.

[31] M. Toplak, R. Močnik, M. Polajnar, Z. Bosnić, L. Carlsson, C. Hasselgren, J. Demšar, S. Boyer, B. Zupan, J. Stålring, Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models, J. Chem. Inf. Model. 54 (2014) 431–441. doi: 10.1021/ci4006595.

[32] L. Shen, D. Cao, Q. Xu, X. Huang, N. Xiao, and Y. Liang, A novel local manifold-ranking based K-NN for modeling the regression between bioactivity and molecular descriptors, Chemom. Intell. Lab. Syst. 151 (2016) 71–77. doi: 10.1016/j.chemolab.2015.12.005.

[33] F. Sahigara, D. Ballabio, R. Todeschini, V. Consonni, Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions, J. Cheminform. 5 (2013) 1–9. doi:10.1186/1758-2946-5-27.

[34] R. Todeschini, D. Ballabio, V. Consonni, F. Sahigara, P. Filzmoser, Locally centred Mahalanobis distance: A new distance measure with salient features towards outlier detection, Anal. Chim. Acta. 787 (2013) 1–9. doi:10.1016/j.aca.2013.04.034.

[35] I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, A. Cherkasov, J. Li, P. Gramatica, K. Hansen, T. Schroeter, K. R. Müller, L. Xi, H. Liu, X. Yao, T. Öberg, F. Hormozdiari, P. Dao, C. Sahinalp, R. Todeschini, P. Polishchuk, A. Artemenko, V. Kuz'Min, T. M. Martin, D. M. Young, D. Fourches, E. Muratov, A. Tropsha, I. Baskin, D. Horvath, G. Marcou, C. Muller, A. Varnek, V. V. Prokopenko, I. V. Tetko, Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set, J. Chem. Inf. Model. 50 (2010) 2094–2111. doi:10.1021/ci100253r.

[36] S. Y. Shin, H. Kim, Extended autoencoder for novelty detection with reconstruction along projection pathway, Appl. Sci. 10 (2020) 4497. doi:10.3390/app10134497.

[37] P. S. Vasafi, O. Paquet-Durand, K. Brettschneider, J. Hinrichs, B. Hitzmann, Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy, J. Food Eng. 299 (2021) 110510. doi: 10.1016/j.jfoodeng.2021.110510.

[38] Z. Boger, Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis, Anal. Chim. Acta. 490 (2003) 31–40. doi:10.1016/S0003-2670(03)00349-0.

[39] V. O. Santos Jr., 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. 547 (2005) 188–196. doi:10.1016/j.aca.2005.05.042.

[40] N. Pasadakis, S. Sourligas, and C. Foteinopoulos, Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks, Fuel. 85, (2006) 1131–1137. doi:10.1016/j.fuel.2005.09.016.

[41] H.A.G. Al-kaf, K.S. Chia, N.A.M. Alduais, A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum, Pet. Sci. Technol. 36 (2018) 411–418. doi:10.1080/10916466.2018.1425717.

[42] K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press London 1st edition (1979).

[43] ASTM E1655-17, Standard Practices for Infrared Multivariate Quantitative Analysis, ASTM Int. (2017) 1–29. doi:10.1520/E1655-17.

[44] H. Moeini, F. M. Torab, Comparing compositional multivariate outliers with autoencoder networks in anomaly detection at Hamich exploration area, east of Iran, J. Geochemical Explor. 180 (2017) 15–23. doi:10.1016/j.gexplo.2017.05.008.

[45] S. Hawkins, H. He, G. Williams, R. Baxter, Outlier detection using replicator neural networks, in: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2000), LNCS 2454, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 170–180. doi:10.1007/3-540-46145-0_17.

[49] D40052-15, Standard Test Method for Density, Relative Density , and API Gravity of Liquids by Digital Density Meter, ASTM Int., (2013) 1–8. doi:10.1520/D4052-18A.2.

[47] R.M. Silverstein, F.X. Webster, D. Kiemle, Spectrometric Identification of Organic Compounds, 7th ed., Jhon Wiley & Sons, Inc, United States of America, 2005. https://books.google.es/books?id=mQ8cAAAAQBAJ.

[48] M. B. Wilk, R. Gnanadesikan, Probability plotting methods for the analysis of data, Biometrika. 55 (1968) 1-17. doi:10.2307/2334448.

IV.7 Supplementary information

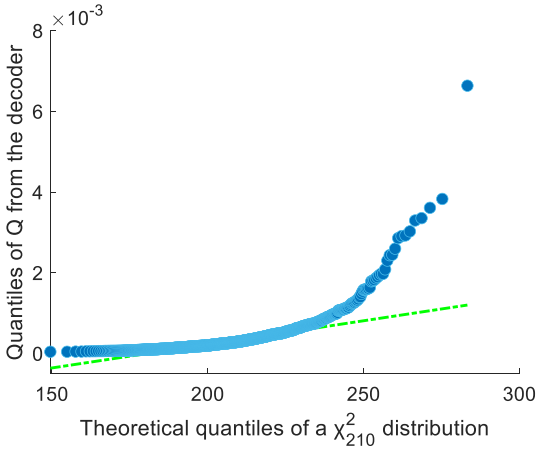


Figure IV-1S. Quantiles of Q from the decoder vs. theoretical quantiles of a χ^2_{210} distribution.

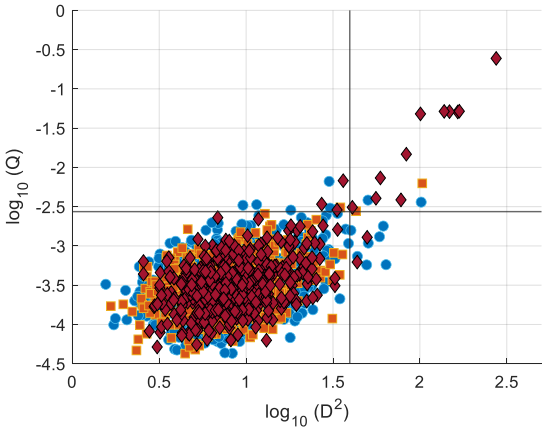


Figure IV-2S. $log_{10}(Q)$ of the decoder vs. $log_{10}(D^2)$ for the training (blue), validation (orange) and test samples (red). The limits of the AD of the regression network are shown. The logarithmic scale was used to facilitate the visualization of the samples around the established limits of the AD.

Chapter V
PLS vs. ANNs calibration models for determining diesel quality parameters

V.1 Introduction

This chapter aims to develop multivariate calibration models for the quantitative determination of the physicochemical properties of desulphurized diesel samples (stream 1) and commercial diesel samples (stream 2) using IR spectroscopy. The physicochemical properties of interest were listed in Chapter I. The calibration models will be based on PLS regression and FFNN. The methodology followed for each quality parameter involves a) selection of the optimal region of the IR spectrum, b) calculation of the models and definition of their limits of applicability (Q and Hotelling's T^2 for PLS and Q and D^2 for FFNN) and c) validation of the models. Additionally, we compare the performance of both models in determining each diesel property and contrast them with performance data from the literature.

V.2 Experimental

V.2.1 IR data and software

The sample set comprised 1736 diesel samples described in section III.1 of Chapter III. All samples were recorded under identical conditions using the procedure outlined in section III.4.1 of Chapter III. From the recorded spectrum, only the spectral ranges of 6021-3275 cm^{-1} and 2760-1800 cm^{-1} were considered for developing the calibration models.

For each stream, 50%, 25%, and 25% of the samples were randomly assigned to the calibration, validation, and test set, respectively, making this division individually for each of the 36 months of diesel production considered in this dataset. To compare the results of PLS and ANN models, the same training, validation, and testing sets were utilized for both models of each property.

PLS models were carried out with Matlab (The MathWorks Inc., Natick, MA, R2022a) and PLS-Toolbox v7 (The Eigen Vector Research, Manson, WA). ANN models were calculated using Matlab's Deep Learning Toolbox.

V.2.2 Selection of the optimal spectral region for the PLS model

Wavenumber selection was performed using a changeable size moving window partial least squares (CSMWPLS) approach. Although this method is computationally intensive compared to other common selection approaches, it was very simple to implement, and the calculation times were reasonable enough for the number of PLS models that had to be tested.

The CSMWPLS approach was applied to two specific regions of interest within the IR spectrum: the NIR-MIR region spanning from 6021 to 3275 cm^{-1} and the MIR region from 2760 to 1800 cm^{-1}. The initial window size was 180 spectral points and was steadily increased to 715 points for the NIR-MIR region and up to 248 points for the MIR region. The starting window size (180) was selected based on the resolution of the spectra and the width of the widest absorption band in the spectrum. For every size window, a PLS model was calculated as the window moved along the spectrum. The PLS models were calculated using mean-centered data, and their number of LVs was chosen as the one with the minimum RMSECV after 10-fold cross-validation. The optimal spectral interval for each IR region of interest was selected as the spectral sub-region corresponding to the size window with the smallest RMSECV value. Subsequently, the PLS model was computed by combining the optimal spectral sub-regions identified: the best one in the NIR-MIR region and the top-performing one in the MIR region. Finally, for each of the two streams, the spectral sub-region or combination of optimal spectral sub-regions with the lowest RMSECV value was chosen for the development of PLS and ANN models.

V.2.3 Calibration models: PLS and ANNs

The PLS models were calculated using mean-centered data in the optimal spectral region selected by CSMWPLS and following the procedure described in the previous section. Meanwhile, the FFNN models were constructed using IR data in the same spectral region selected for the PLS model, employing the procedure described in section IV.3.2 of Chapter IV. The predictive ability of both PLS and ANN models was evaluated as the mean square error of prediction (RMSEP) of the test set for each stream.

The applicability domain of the PLS model was established based on Hotelling's T^2 and Q statistics, as commented in Chapter II. The limits of applicability of the FFNN model were set as the 0.99 quantiles of the squared Mahalanobis distance (D^2) and the spectral residuals (Q) obtained by the autoencoder and the decoder networks, as detailed in Chapter IV. The architecture of both the autoencoder and decoder networks was tested following the procedure described in section IV.3.2 of Chapter IV.

Once the limits of applicability have been defined, a new spectrum is projected onto the calibration model space. Hotelling's T^2 and Q-residuals (for the PLS model) or squared Mahalanobis distance and Q-residuals (for the FFNN model) are calculated and compared to these limits. A sample whose statistics values fall within the limits of the applicability is regarded to be similar to the calibration and validation samples, and its prediction by the model is deemed reliable. Conversely, the prediction of a sample whose

spectrum is beyond the limits of applicability should not be trusted and must be analyzed with the reference method.

The tolerance limits admitted by the models were as follows: density (1 kg/m^3), T65% (2.58 ºC), T85% (3.40 ºC), T95% (6.46 ºC), flash point (3.57 ºC), cloud point (2.36 ºC), CFPP (2.59 ºC), sulfur (1.73 mg/kg) and FAME content (0.66 % v/v), viscosity (0.022 mm^2/s), cetane number (1.08). To evaluate the agreement between the predicted values from the model and those determined by the reference method and also consider whether the model was valid or not for routine analysis, eq. II.13, detailed in Chapter II, was applied. According to the ASTM-E-1655 standard, the models were considered valid for routine analysis if more than 95% of samples exhibited prediction errors within the tolerance limits admitted by the reference method.

V.3 Results and discussion

V.3.1 PLS and ANN models for predicting density

For selecting the optimal spectral range, Figure V-1 and Figure V-2 show the cross-validation error of PLS models constructed for the density of samples from streams 1 and 2 as a function of the starting point and length of the spectral region. In the NIR-MIR region (Figure V-1 (left) and Figure V-2 (left)), the error decreased as the starting point shifted towards smaller wavenumbers and the region length increased. The minimum error for stream 1 was 0.78 kg/m^3, corresponding to the interval 5538.56-3336.25 cm^{-1} (571 points) (Figure V-3). For stream 2, the minimum error was 0.82 kg/m^3, corresponding to the interval 5168.29-3270.68 cm^{-1} (492 points) (Figure V-4). Both spectral regions correspond to vibrational modes of the C-H bond stretching of methyl and methylene group and a combination bands of C-H bond stretching and C=O bond stretching (see Table III-2 of Chapter III).

In the MIR region (Figure V-1 (right) and Figure V-2 (right)), the error for all PLS models is lower than in the NIR-MIR region. As the starting point of the range shifted towards lower wavenumbers, the error decreased. The smallest error for stream 1 was 0.48 ($\mathrm{kg/m^3}$), corresponding to the interval 2518.58-1824.33 cm^{-1} (180 points) (Figure V-3). For stream 2, the smallest error was 0.56 ($\mathrm{kg/m^3}$) corresponding to the interval 2742.28-1801.19 cm^{-1} (244 points) (Figure V-4). Both regions are characteristic of the fundamental vibrations of the C-H, C≡C, C=O, and S-H bonds (see Table III-2 of Chapter III).

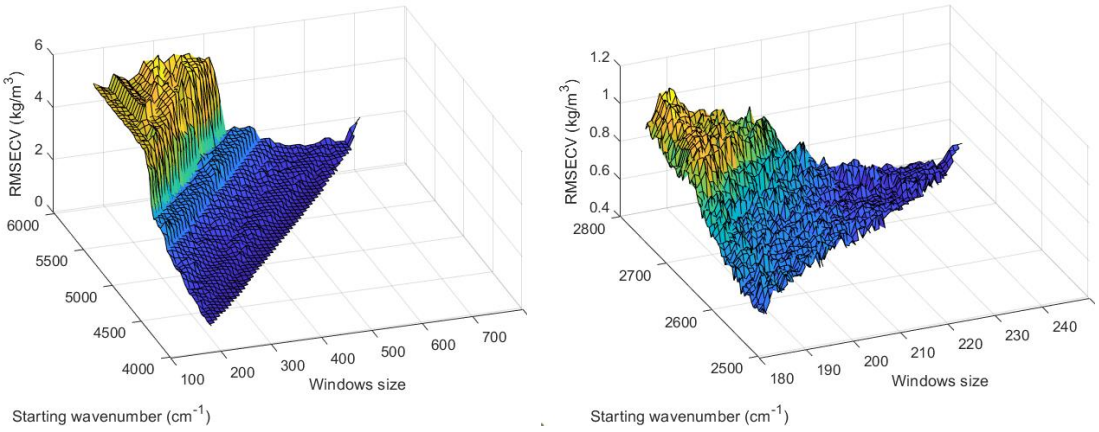


Figure V-1. Cross-validation error of each PLS model for density prediction of desulfurized samples during the CSMWPLS procedure as a function of the starting wavenumber of the spectral window and the window size in the NIR-MIR (left) and MIR (right) region.

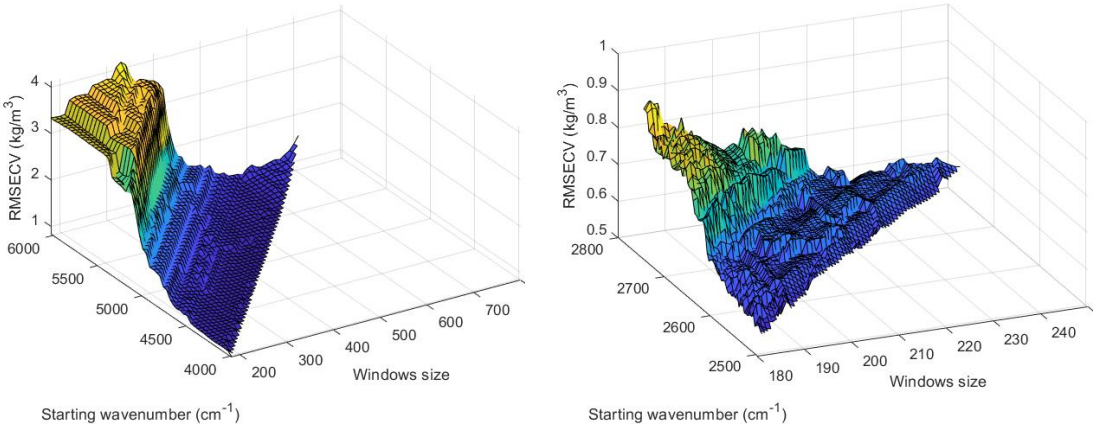


Figure V-2. Cross-validation error of each PLS model for density prediction of commercial samples during the CSMWPLS procedure as a function of the starting wavenumber of the spectral window and the window size in the NIR-MIR (left) and MIR (right) regions.

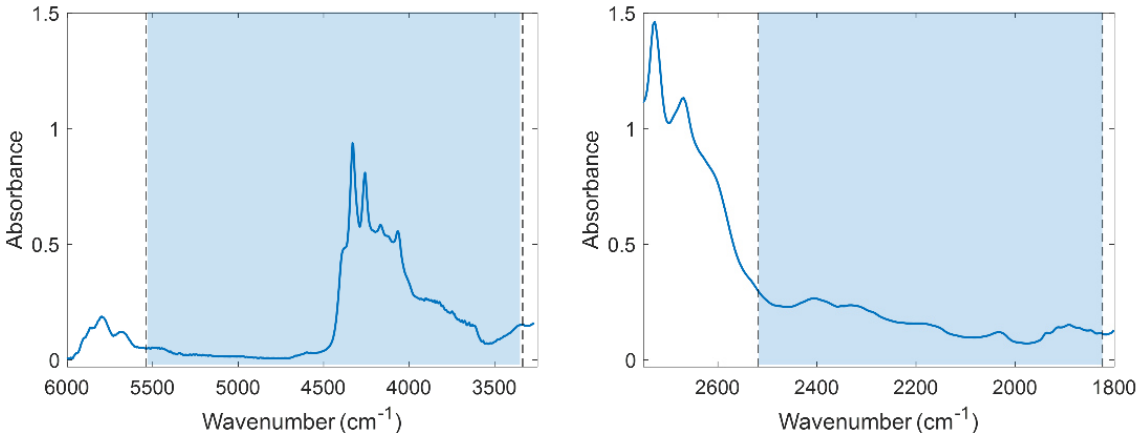


Figure V-3. Optimal spectral range in the NIR-MIR (left) and MIR (right) regions for the PLS model of desulfurized samples.

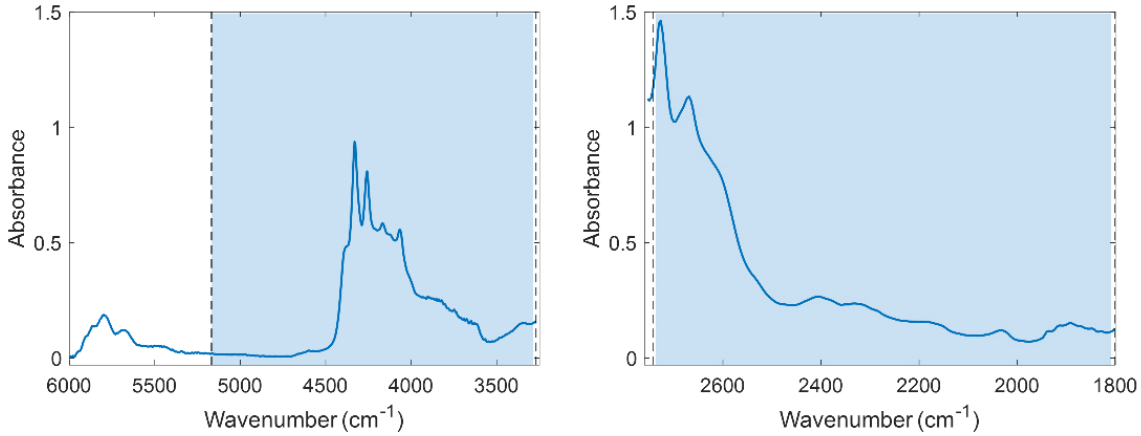


Figure V-4. Optimal spectral range in the NIR-MIR (left) and MIR (right) regions for the PLS model of commercial samples.

Table V-1 shows the main results of the best-performing PLS models developed for samples from each stream in the best NIR sub-region, the best MIR sub-region, and the union of both selected sub-regions. As can be seen, the RMSECV of the PLS model for samples from stream 1 did not improve by joining the spectral sub-regions, but it improved for samples from stream 2. Hence, the PLS model for samples from stream 1 was based on the optimal MIR sub-region (2518.58-1824.33 cm^{-1}) selected, whereas the PLS model for samples from stream 2 was based on the joint NIR-MIR sub-regions (5168.29-3270.68 and 2742.28-1801.19 cm^{-1}).

Table V-1. Characteristics of the PLS models in the optimal spectral subregion NIR-MIR, MIR, and the combination NIR-MIR and MIR intervals.

	PLS	
	Stream 1	**Stream 2**
Selected region (NIR-MIR) $\mathbf{cm^{-1}}$	5538.56-3336.25	5168.29-3270.68
LVs	10	10
RMSECV	0.78	0.83
Selected region (MIR) $\mathbf{cm^{-1}}$	2518.58-1824.33	2742.28-1801.19
LVs	10	13
RMSECV	0.48	0.56
Region (NIR-MIR) $\mathbf{cm^{-1}}$	5538.56-3336.25 and 2518.58-1824.33	5168.29-3270.68 and 2742.28-1801.19
LVs	12	14
RMSECV	0.58	0.48

Figure V-5 shows the applicability domain of the PLS model for samples from streams 1 and 2, expressed for each statistic, corresponding to significance levels of

95% and 99%. As can be seen, at a significance level of 95%, only a few calibration samples from stream 2 had high Q residuals and Hotelling's T^2 values. Nevertheless, these samples that remained very close to the limits at this significance level had low prediction errors, and hence, they were kept in the model. At a significance level of 99%, no calibration sample from streams 1 and 2 was excluded from the analysis. Those samples from the test set of each stream that exceeded, Q_{lim} and T^2_{lim} with high prediction errors were removed. Some of these removed samples had density values outside the range of the calibration model for each stream.

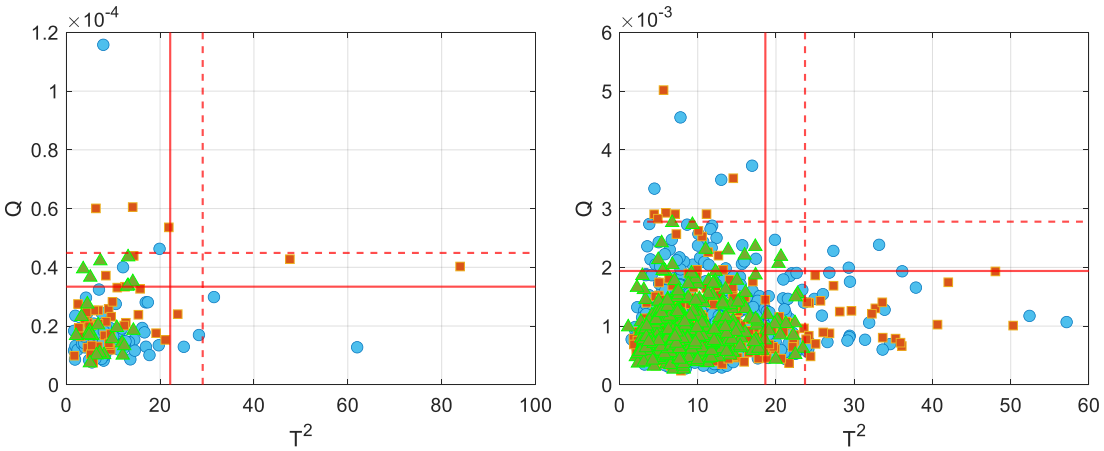


Figure V-5. Q residuals vs. Hotelling's T^2 for stream 1 (left) and stream 2 (right) PLS models. Solid and dashed lines represent the limit of Hotelling's T^2 and Q statistics at 95% and 99% confidence, respectively. Calibration samples (blue), validation samples (orange), and test samples (green).

Figure V-6 shows the experimental density values versus those estimated by the PLS models for samples from each stream. The determination coefficient (R^2) of the linear fit between the reference density and the predictions of each PLS model was 0.99. The RMSE values for the calibration, validation, and test samples from stream 1 were 0.34 kg/m^3, 0.61 kg/m^{3,} and 0.73 kg/m^3, respectively. Meanwhile, the corresponding values for the calibration, validation, and test samples from stream 2 were 0.40 kg/m^3, 0.43 kg/m^{3,} and 0.46 kg/m^3, respectively. The predictive ability of the PLS model for stream 2 was better than that for stream 1.

Figure V-7 shows the prediction error of the test samples from streams 1 and 2 in temporal order and the allowed tolerance limit for density. It is observed that there is no systematic temporal trend in the prediction error. Based on the ASTM-E-1655, only the IR/PLS model for stream 2 was considered to give estimated values that agree with the reference method.

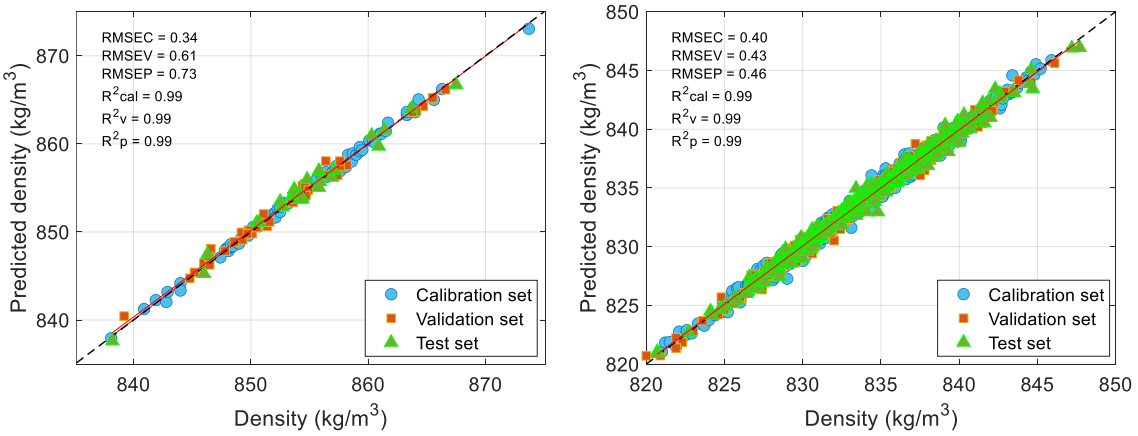


Figure V-6. Predicted density vs. measured value for stream 1 (left) and stream 2 (right) with PLS models.

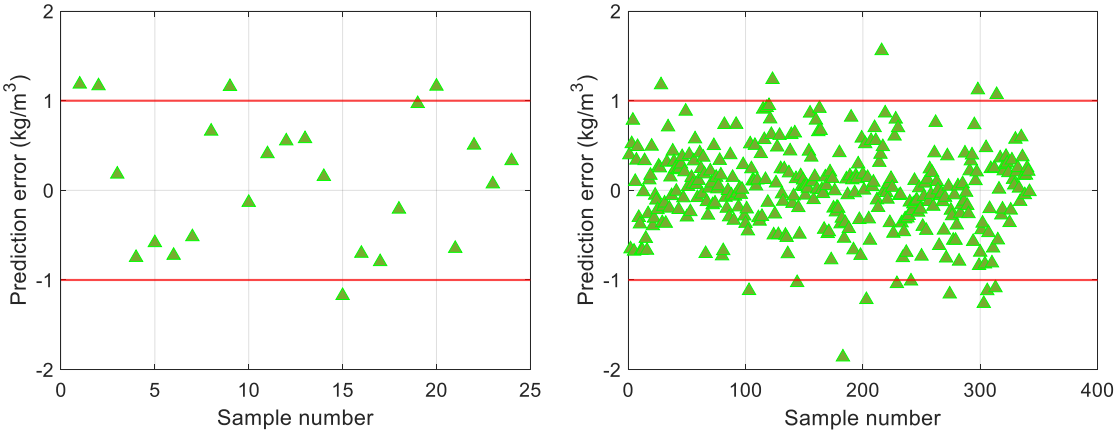


Figure V-7. Prediction error of the test set for streams 1 (left) and 2 (right) of PLS models in the temporal order. Red lines represent the tolerance limits admitted.

Figure V-8 shows the experimental density values versus those predicted by the FFNN models for the training, validation, and test samples from streams 1 and 2. The RMSE values for the calibration, validation, and test samples from stream 1 were 0.39 kg/m^3, 0.89 kg/m^3, and 0.85 kg/m^3, respectively. The R^2 of the linear fit between the reference density and the predictions was 0.99, 0.98, and 0.98 for the training, validation, and test sets of the same stream, respectively. For samples from stream 2, R^2 was also high (0.99), and the RMSE was 0.25 kg/m^3, 0.44 kg/m^3, and 0.42 kg/m^3 for the calibration, validation, and test sets, respectively. Similar to the PLS models, the predictive ability of the FFNN models for samples from stream 1 was slightly worse than that obtained for samples from stream 2, probably due to the small set used for the training network.

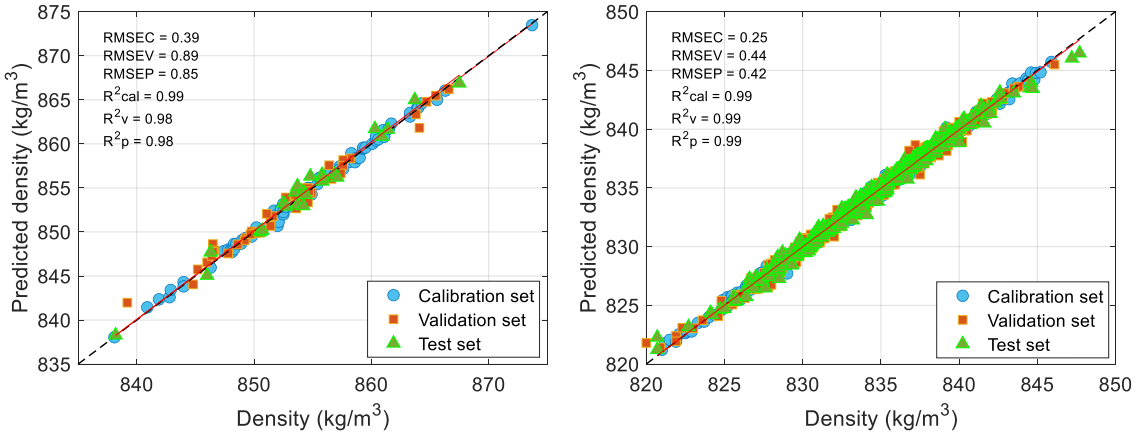


Figure V-8. Predicted density vs measured value for stream 1 (left) and stream 2 (right) for FFNN models.

Similar to the approach described in section IV.4.1, different autoencoder architectures with one or three hidden layers were tested to reproduce the training spectra of each FFNN model. The selected autoencoder for stream 1 had 10, 5, and 10 neurons in the first, middle (bottleneck), and third hidden layers, respectively. For stream 2 the autoencoder architecture had 15, 10, and 15 neurons in the first, middle (bottleneck), and third hidden layers, respectively. Figure V-9 and Figure V-10 show the reconstructed spectra $\hat{\mathbf{x}}$ and the spectral residuals of the calibration and validation samples of streams 1 and 2, respectively. For both the calibration and validation sets of streams 1 and 2, the correlation coefficient between each spectrum and the reconstructed spectrum ranged from 0.99 to 1. The Q values were calculated from the spectral residuals of the autoencoder and the squared Mahalanobis distance of calibration, validation, and test spectra was calculated from the activations of the hidden layer of the regression network for two streams.

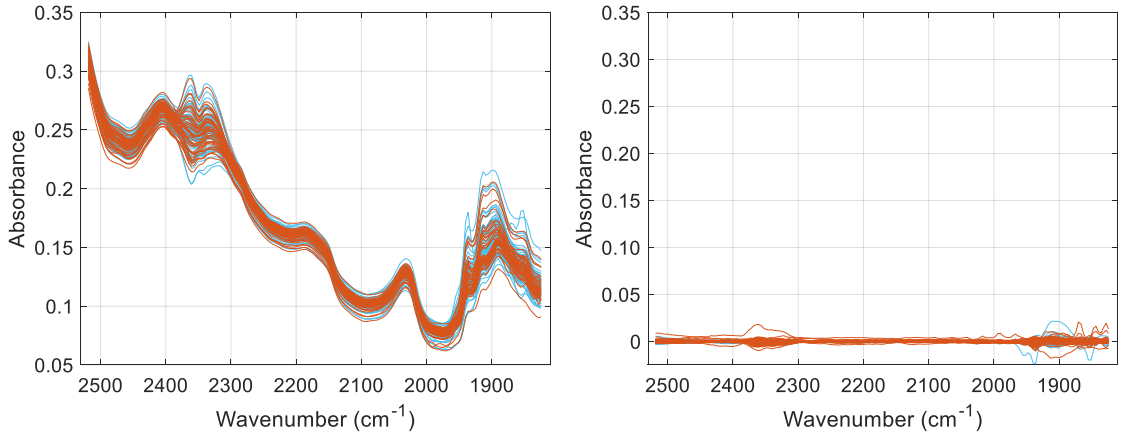


Figure V-9. Reconstructed training and validation spectra of samples from stream 1 (left) and spectral residuals (right) from the autoencoder. Brown and blue lines represent the training and validation spectra, respectively.

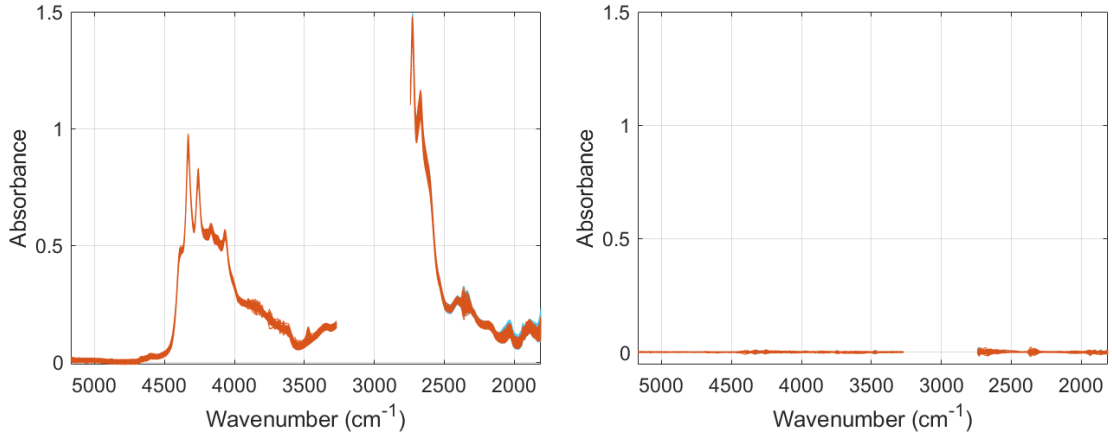


Figure V-10. Reconstructed training and validation spectra of samples from stream 2 (left) and spectral residuals (right) from the autoencoder. Brown and blue lines represent the training and validation spectra, respectively.

Figure V-11 shows the limits of the applicability domain of the regression network defined by the squared Mahalanobis distance and the spectral residuals for streams 1 and 2. As expected, the samples showed different trends in both applicability domains with respect to those observed for each PLS model, and some samples from both streams were outside the AD limits. In stream 1, the only calibration sample with Q and D^2 values larger than Q_{lim} and D^2_{lim} corresponds to the sample with the highest density. In stream 2, some samples of the validation set are also outside the established applicability domain. In both cases, the network did not show prediction errors greater than the admitted tolerance limit for these samples. The AD limits detected no discordant spectra in the test set with a Q and D^2 values larger than Q_{lim} and D^2_{lim}.

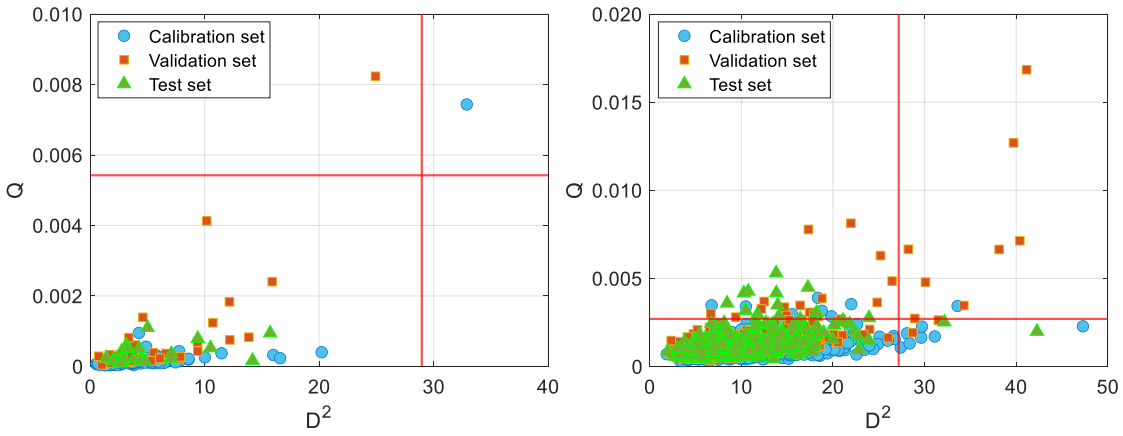


Figure V-11. Q of the autoencoder vs. D^2 for the training (blue), validation (orange), and test samples (green) of stream 1(left) and stream 2 (right). The limits of the applicability domain of the regression network are shown.

As an alternative to the autoencoder, which is trained independently on the regression network, the spectral residual limit of the AD was also calculated from the decoder part of an autoencoder, calculated from the activations of the hidden layer of

the regression network. The simplest network with one hidden layer and 15 neurons was selected as it provided the lowest RMSE for the validation set. Figure V-12 and Figure V-13 show the spectra estimated by the decoder for the training and validation sets and their corresponding spectral residuals for samples from streams 1 and 2, respectively. For both the calibration and validation sets, the correlation coefficient between each spectrum and the estimated spectrum ranged between 0.9985 and 1 for both stream 1 and stream 2. As expected, the spectral residuals are larger than the residuals produced by the autoencoder.

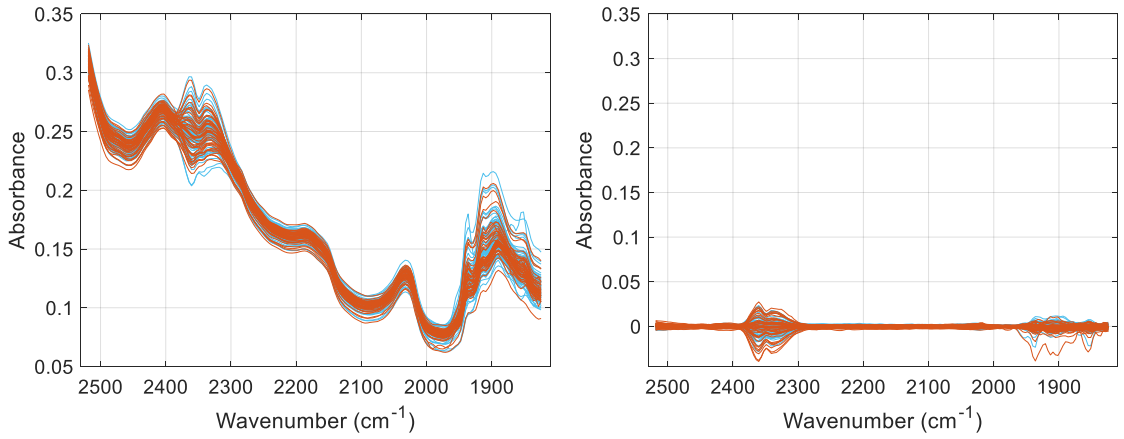


Figure V-12. Reconstructed training and validation spectra of stream 1(left) and spectral residuals (right) from the decoder. Brown and blue lines represent the training and validation spectra, respectively.

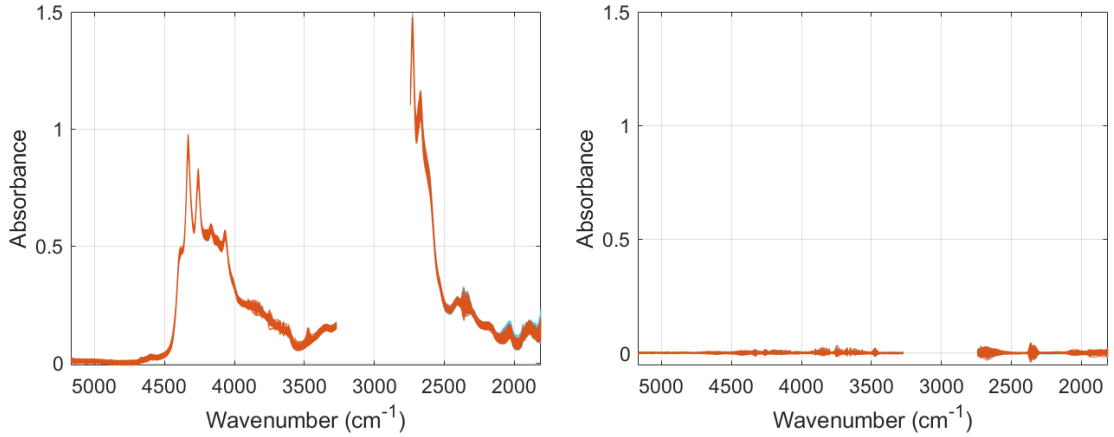


Figure V-13. Reconstructed training and validation spectra of stream 2(left) and spectral residuals (right) from the decoder. Brown and blue lines represent the training and validation spectra, respectively.

Figure V-14 shows the AD limits of the regression network defined by the squared Mahalanobis distance and the spectral residuals of the autoencoder and decoder, respectively for the two streams. As it was foreseen, for both streams the decoder reconstructs the spectra worse than the autoencoder, resulting in larger residuals and a larger Q_{lim}. The AD limits established from the squared Mahalanobis distance and the

autoencoder also detected no discordant spectra in the test set with a Q and D^2 values larger than Q_{lim} and D^2_{lim}.

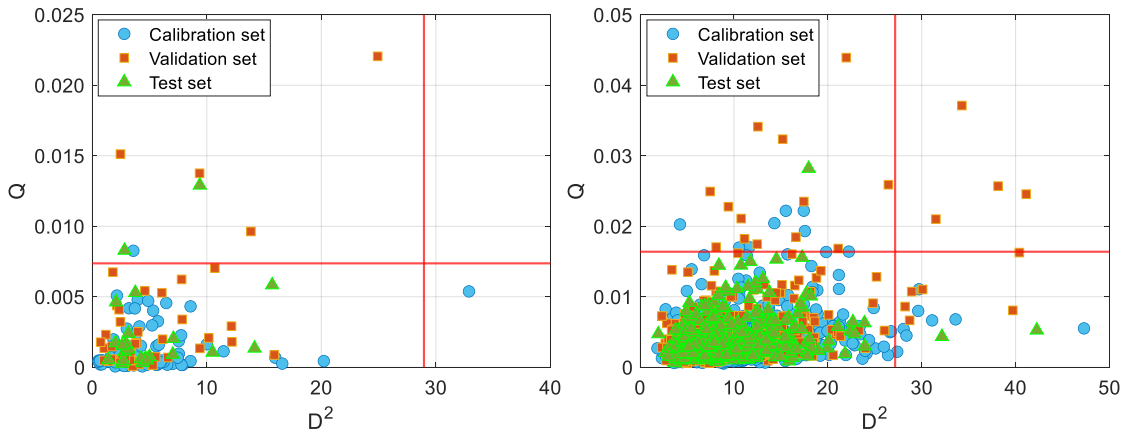


Figure V-14. Q of the decoder vs. D^2 for the training (blue), validation (orange), and test samples (green). The limits of the applicability domain of the regression network are shown.

Figure V-15 shows the prediction error of the density for each sample $e_i = |y_i - \hat{y}_i|$ of the test set from streams 1 and 2 in temporal order. As for the results of PLS (Figure V-7), the prediction error is random, without a temporal trend. Besides, the IR/ANN model for stream 2 was also considered to give predictions that agree with the reference method. In both streams, some samples tested that were within the Q_{lim} and D^2_{lim} fell beyond the allowed tolerance limit for density, but their prediction error was low.

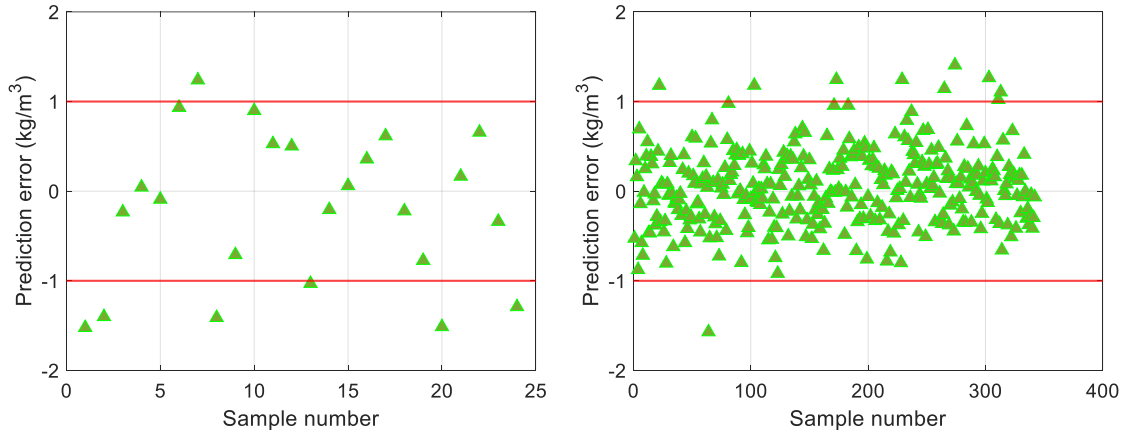


Figure V-15. Prediction error of the test set for streams 1 (left) and 2 (right) for FFNN models in temporal order.

Method comparison

The calibration, validation, and test results of the PLS and ANN models are compared in Table V-2. Only five nodes were needed in the network, while PLS required

10 LVs for stream 1. The network for samples from stream 2 also needed a smaller number of nodes (10) than the number of LVs in the PLS model (14). Taking into account the prediction errors of models for the test sets of streams 1 and 2, the PLS and ANN models of stream 2 yielded a lower RMSEP for the NIR-MIR method. The network method provided a slightly lower error than the PLS model for the test set of stream 2.

Table V-2. Characteristics of the best PLS and FFNN models for predicting density in diesel samples of streams 1 and 2.

	PLS		FFNN	
	stream 1	stream 2	stream 1	stream 2
LVs and nodes	10	14	5	10
Cumulative Variance X, Y (%)	98.84%, 99.72%	98.97%, 99.17%	-	-
RMSEC	0.34	0.40	0.39	0.25
R^2c	0.99	0.99	0.99	0.99
RMSEV	0.61	0.43	0.89	0.44
R^2v	0.99	0.99	0.98	0.99
RMSEP	0.73	0.46	0.85	0.42
R^2p	0.99	0.98	0.98	0.99

Comparison with available literature data

Figure V-16 compares the above results in terms of RMSEP for samples from both streams using PLS and ANN models with those from references found in the literature review (Table 4 in Chapter I). Our PLS calibration models for samples from streams 1 and 2 yielded similar results to those of Bezerra de Lira et al. (2010) [1], Ferrão et al. (2010) [2], Marinović et al. (2012) [3], and Nespeca et al. (2018) [4], the RMSEP value for stream 1 always being slightly higher. Except for the result reported by Santos Jr. et al. (2005) [5], our PLS models for both streams obtained better results than those of the remaining works. Our ANN models for streams 1 and 2 yielded similar results to those of Al-kaf et al. (2018) [6] and Santos Jr. et al. (2005) [5], respectively. The behavior of the predictive abilities of models found in the literature and those developed in the current study are primarily attributed to the variations in diesel feedstock, the disparity in the number of samples in the datasets, and the differences in the ranges of the density values. Comparatively with the reviewed studies, the added value of our model results for diesel samples from stream 2 lies in the fact that they are based on a broader set of samples that covers a lot of variability in the production process.

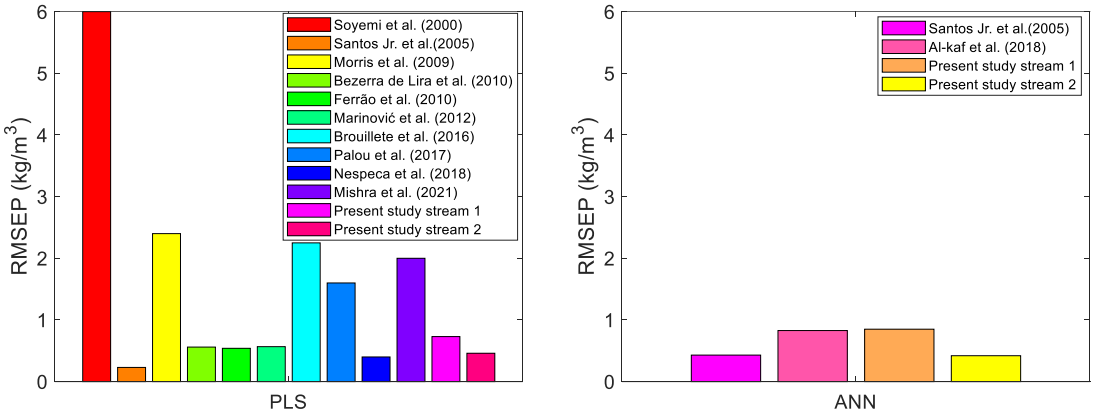


Figure V-16. Comparison of our RMSEP values of PLS (left) and ANN (right) for prediction of density in diesel-diesel/biodiesel blends from its IR spectra with those from the literature.

Partial conclusions

The following partial conclusions can be drawn:

1. The optimal spectral region to develop the PLS model was the MIR region between 2518.58-1824.33 cm^{-1} for samples from stream 1, and the joining NIR-MIR and MIR subregions between 5168.29-3270.68 and 2742.28-1801.19 cm^{-1} for samples from stream 2.
2. The RMSEPs values of both calibration approaches, PLS and FFNN, on the same test set were comparable to those reported in the revised literature. RMSEPs values of PLS models for samples from streams 1 and 2 were 0.73 kg/m^3 and 0.46 kg/m^3, respectively, while for FFNN models, the corresponding values were 0.85 kg/m^3 and 0.42 kg/m^3. Comparatively, the predictive ability of the FFNN model developed for each stream was similar to that obtained by the PLS model, being slightly inferior for stream 1.
3. The limits of applicability of both PLS and FFNN models are useful to detect samples that might not be similar to the samples used to establish the model. The Q-residuals for the FFNN model calculated from a decoder were higher than those of the autoencoder. Hence, the decoder approach was considered to define the limit of Q for other properties of interest.
4. The percentage of the test samples whose prediction error falls outside the tolerance limits admitted by the reference method were 20.8% and 4.4% for the PLS models for samples from streams 1 and 2, respectively, and 25% and 2.9% for FFNN models. Therefore, based on ASTM-E-1655, we recommend only the

IR/PLS and IR/ANN models of stream 2 for practical implementation in routine diesel analysis.

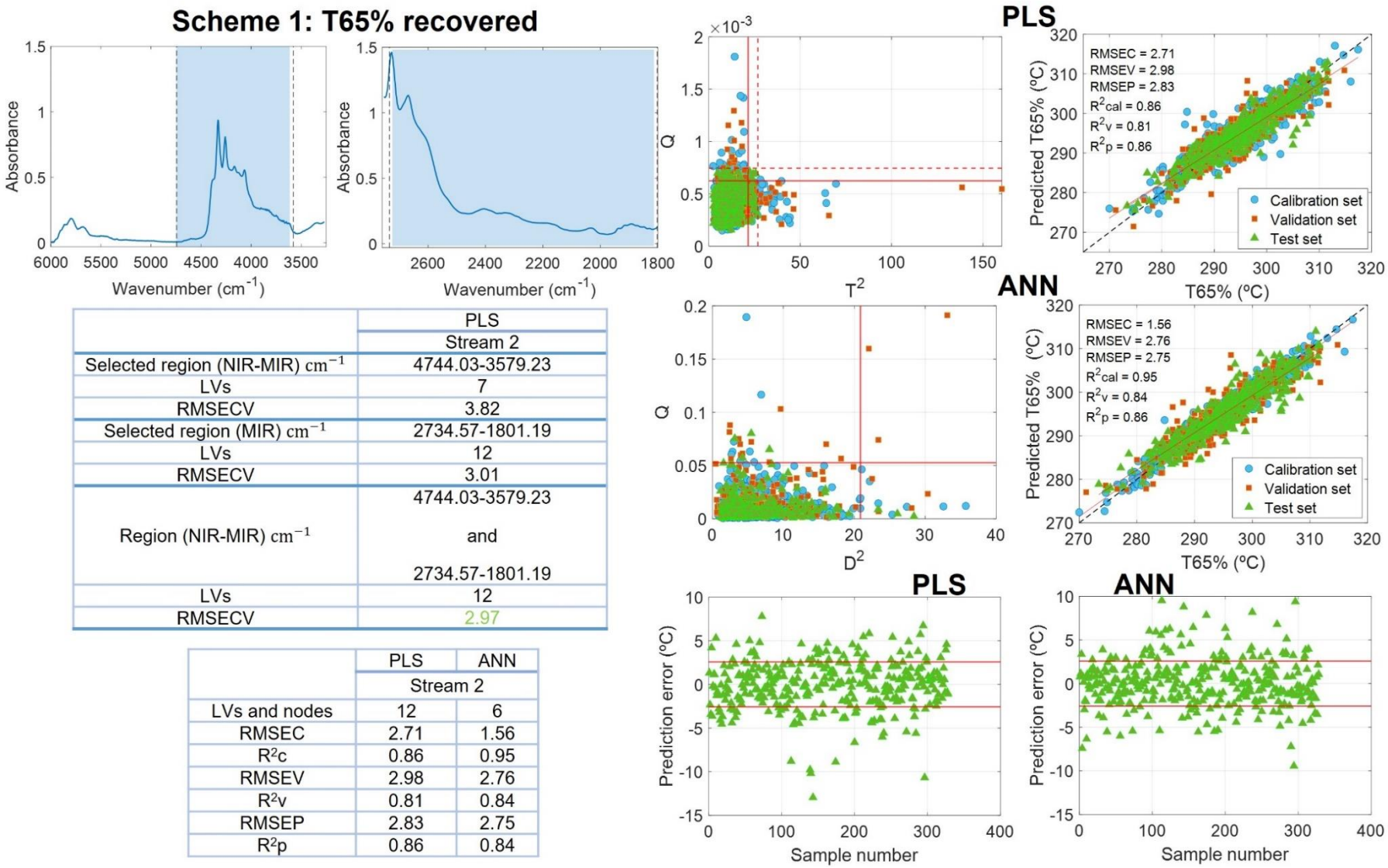
V.3.2 Results of PLS and ANN models of the remaining properties

Schemes 1 to 15 show in the condensed format the results of the variable selection procedure, the applicability domain corresponding to each model, the predicted versus measured values, and the prediction error of the test set in temporal order, corresponding to PLS and ANN models for predicting the properties of interest in samples from each stream. In addition, the RMSEP values of PLS and ANNs for properties of interest (Figure V-17) are compared with those from the literature (Table I-5 in Chapter I). T65% was not included in the comparison as only SEP values were available in the literature.

The comparative analysis of our results with those reported in the literature (Figure V-17*)* revealed the following aspects:

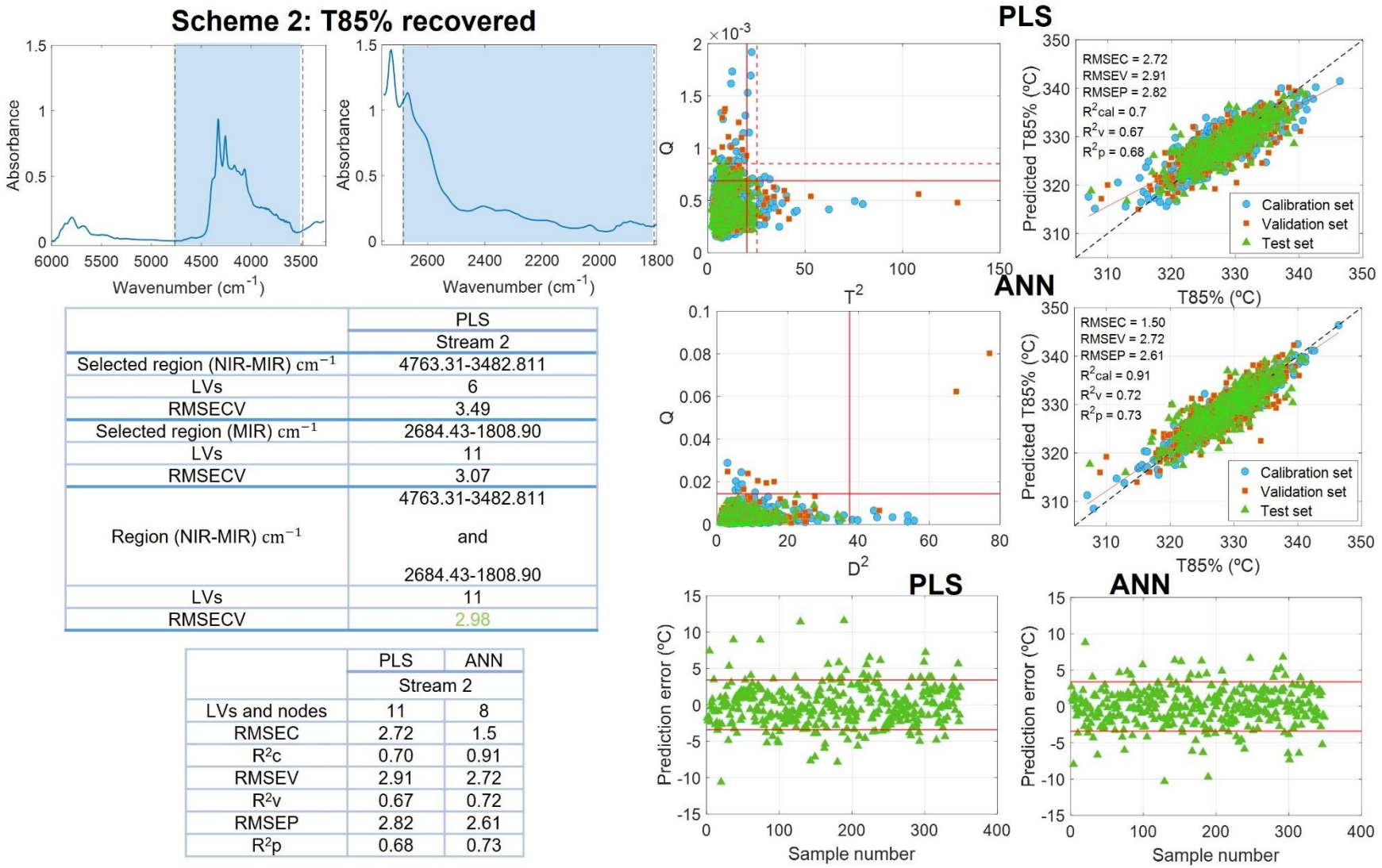
- For T85%, our PLS calibration model yielded an RMSEP value similar to those obtained in [5,7], albeit slightly lower. In addition, the corresponding value for the ANN model was comparable to that reported by Santos Jr. et al. (2005) [5].
- For T95%, our PLS models exhibited higher predictive abilities for samples from stream 2 and lower abilities for samples from stream 1 compared to those reported in Palou et al. [8].
- For the cetane number of samples from stream 1, the RMSEP value by the PLS model was similar to those reported by [4,9], whereas for samples from stream 2 the RMSEP was similar to that obtained by Brouillete et al. [10]. Furthermore, the RMSEP values of the corresponding ANN models were lower than that of Al-kaf et al. [6].
- For viscosity, both PLS and ANN model achieved the best predictive ability compared to references [3,5,6,10–14].
- For flash point, the RMSEP obtained by PLS for samples from stream 1 was similar to those achieved by Brouillete et al. [10], while for samples from stream 2 was similar to those obtained by Nespeca et al. [4].
- For FAME content, the RMSEP value from the PLS model was equal to the one reported by Nespeca et al. [4].
- For cloud point, the RMSEP obtained for PLS models of samples from stream 1 and 2 were higher than that obtained by Palou et al. [8] and significantly lower than that obtained by Brouillete et al. [10], respectively.

- For CFPP, the ability predictive of the PLS model was inferior to that obtained by Hradecká et al.[14].
- For sulfur content, the RMSEP of our PLS model for samples from stream 2 was slightly lower but comparable to that obtained by Palou et al. [8], whereas the corresponding value for samples from stream 1 was higher than those obtained by [8,14–16].



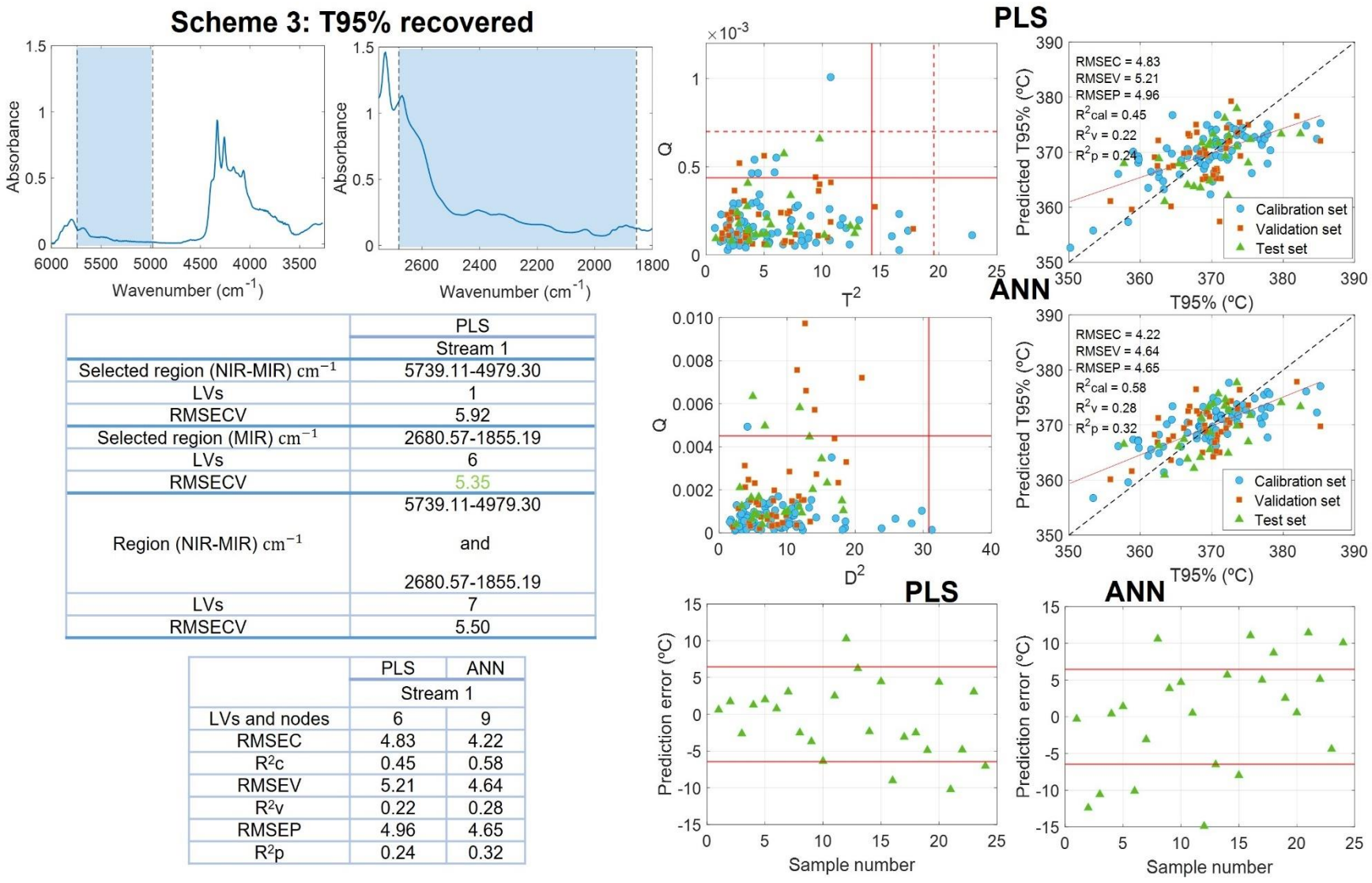
	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4744.03-3579.23
LVs	7
RMSECV	3.82
Selected region (MIR) cm^{-1}	2734.57-1801.19
LVs	12
RMSECV	3.01
Region (NIR-MIR) cm^{-1}	4744.03-3579.23 and 2734.57-1801.19
LVs	12
RMSECV	2.97

	PLS	ANN
	Stream 2	
LVs and nodes	12	6
RMSEC	2.71	1.56
R^2c	0.86	0.95
RMSEV	2.98	2.76
R^2v	0.81	0.84
RMSEP	2.83	2.75
R^2p	0.86	0.84



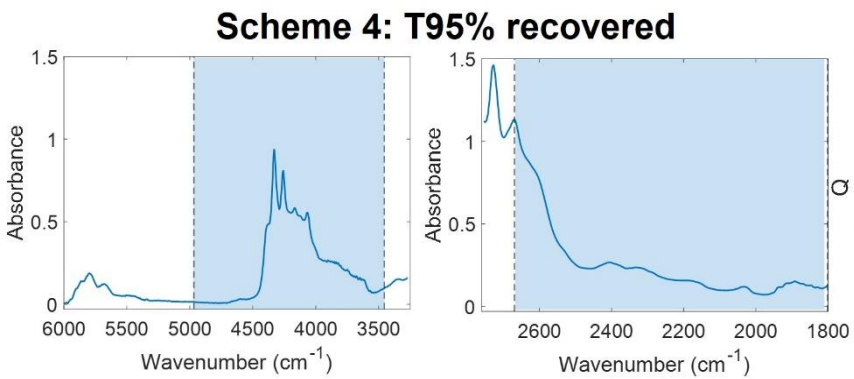
	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4763.31-3482.811
LVs	6
RMSECV	3.49
Selected region (MIR) cm^{-1}	2684.43-1808.90
LVs	11
RMSECV	3.07
Region (NIR-MIR) cm^{-1}	4763.31-3482.811 and 2684.43-1808.90
LVs	11
RMSECV	2.98

	PLS	ANN
	Stream 2	
LVs and nodes	11	8
RMSEC	2.72	1.5
R^2c	0.70	0.91
RMSEV	2.91	2.72
R^2v	0.67	0.72
RMSEP	2.82	2.61
R^2p	0.68	0.73



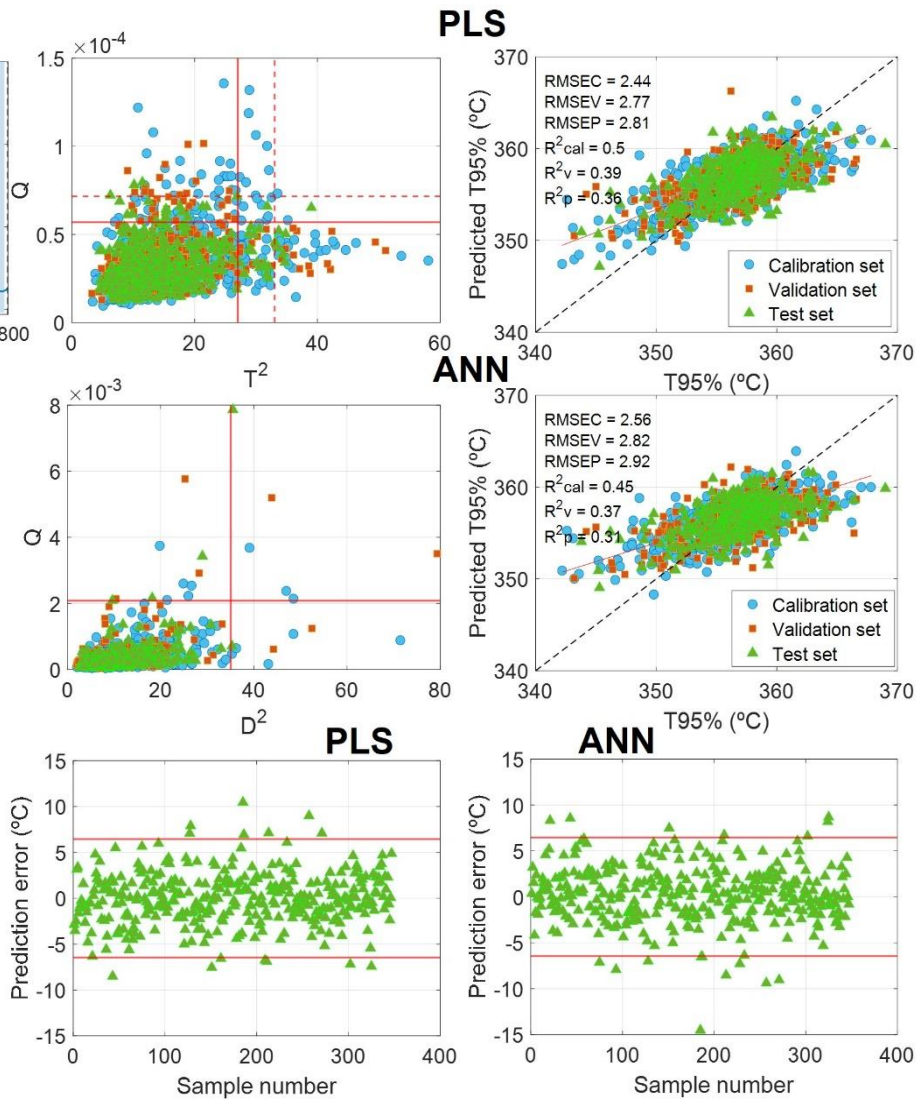
	PLS
	Stream 1
Selected region (NIR-MIR) cm^{-1}	5739.11-4979.30
LVs	1
RMSECV	5.92
Selected region (MIR) cm^{-1}	2680.57-1855.19
LVs	6
RMSECV	5.35
Region (NIR-MIR) cm^{-1}	5739.11-4979.30 and 2680.57-1855.19
LVs	7
RMSECV	5.50

	PLS	ANN
	Stream 1	
LVs and nodes	6	9
RMSEC	4.83	4.22
R^2c	0.45	0.58
RMSEV	5.21	4.64
R^2v	0.22	0.28
RMSEP	4.96	4.65
R^2p	0.24	0.32

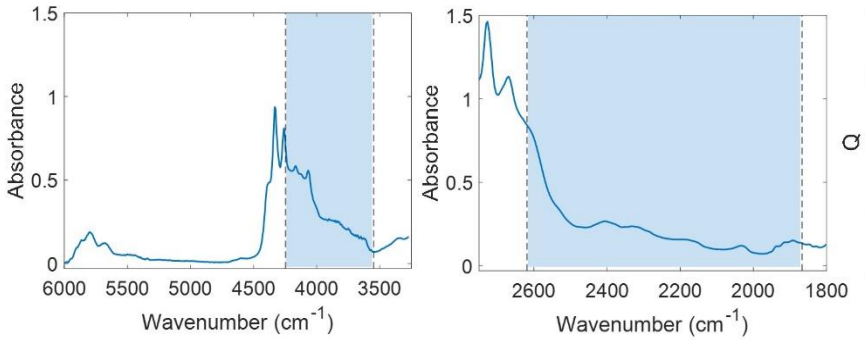


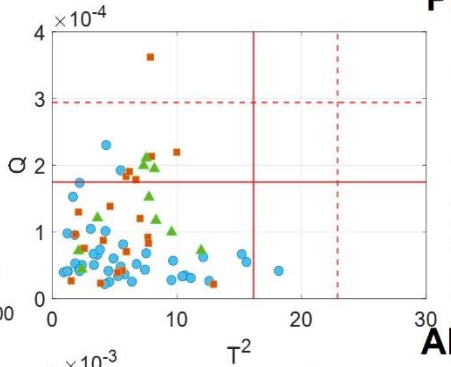
	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4967.73-3455.81
LVs	8
RMSECV	3.10
Selected region (MIR) cm^{-1}	2669.0-1801.19
LVs	12
RMSECV	2.89
Region (NIR-MIR) cm^{-1}	4967.73-3455.81 and 2669.0-1801.19
LVs	11
RMSECV	3.01

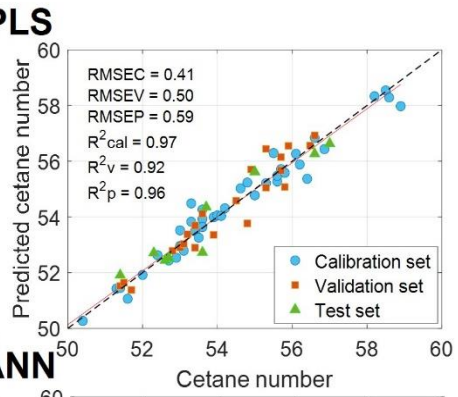
	PLS	ANN
	Stream 2	
LVs and nodes	16	11
RMSEC	2.44	2.56
R^2c	0.50	0.45
RMSEV	2.77	2.82
R^2v	0.39	0.37
RMSEP	2.81	2.92
R^2p	0.36	0.31

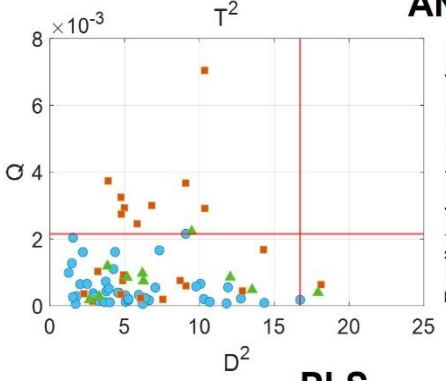


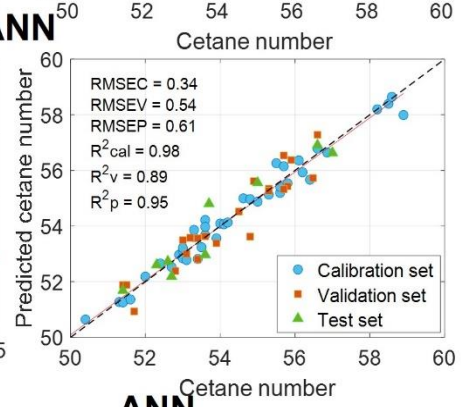
Scheme 5: Cetane number





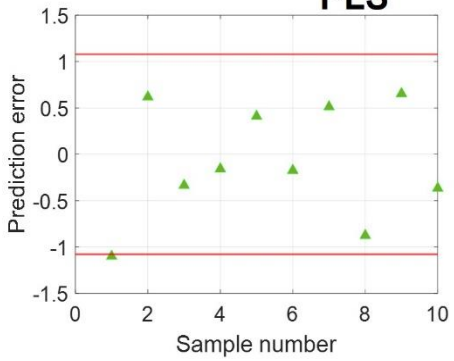


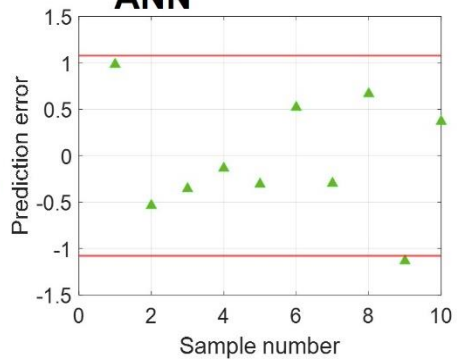




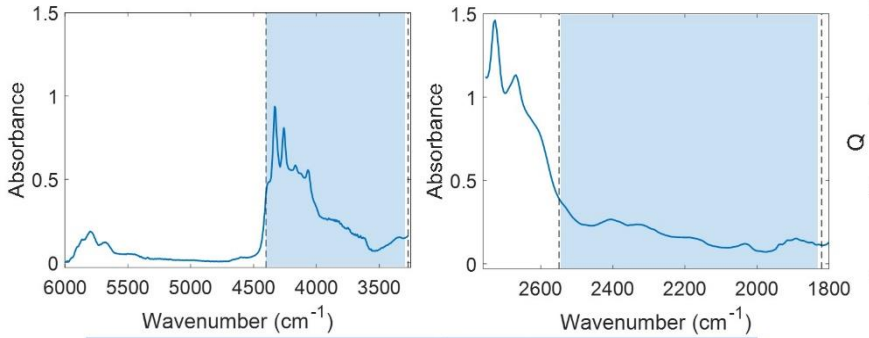
	PLS
	Stream 1
Selected region (NIR-MIR) cm^{-1}	4246.48-3548.38
LVs	6
RMSECV	0.75
Selected region (MIR) cm^{-1}	2618.86-1866.76
LVs	6
RMSECV	0.57
Region (NIR-MIR) cm^{-1}	4246.48-3548.38 and 2618.86-1866.76
LVs	6
RMSECV	0.65

	PLS	ANN
	Stream 1	
LVs and nodes	6	6
RMSEC	0.41	0.34
R^2c	0.97	0.98
RMSEV	0.50	0.54
R^2v	0.92	0.89
RMSEP	0.59	0.61
R^2p	0.96	0.95



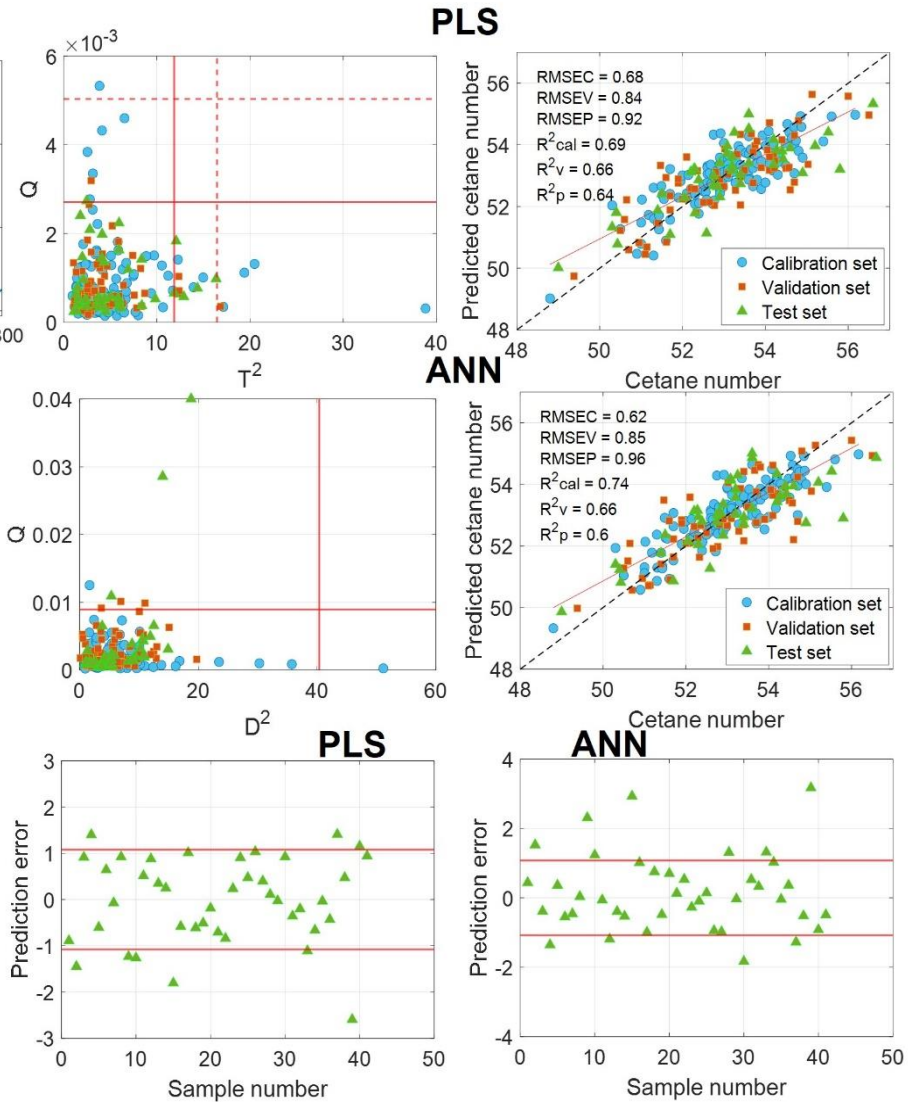


Scheme 6: Cetane number

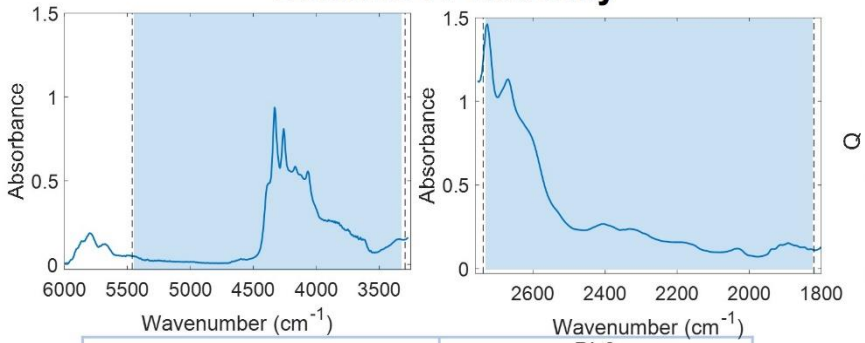


	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4400.76-3270.68
LVs	5
RMSECV	0.74
Selected region (MIR) cm^{-1}	2549.43-1820.47
LVs	5
RMSECV	0.78
Region (NIR-MIR) cm^{-1}	4400.76-3270.68 and 2549.43-1820.47
LVs	6
RMSECV	0.75

	PLS	ANN
	Stream 2	
LVs and nodes	5	6
RMSEC	0.68	0.62
R^2c	0.69	0.74
RMSEV	0.84	0.85
R^2v	0.66	0.66
RMSEP	0.92	0.96
R^2p	0.64	0.60

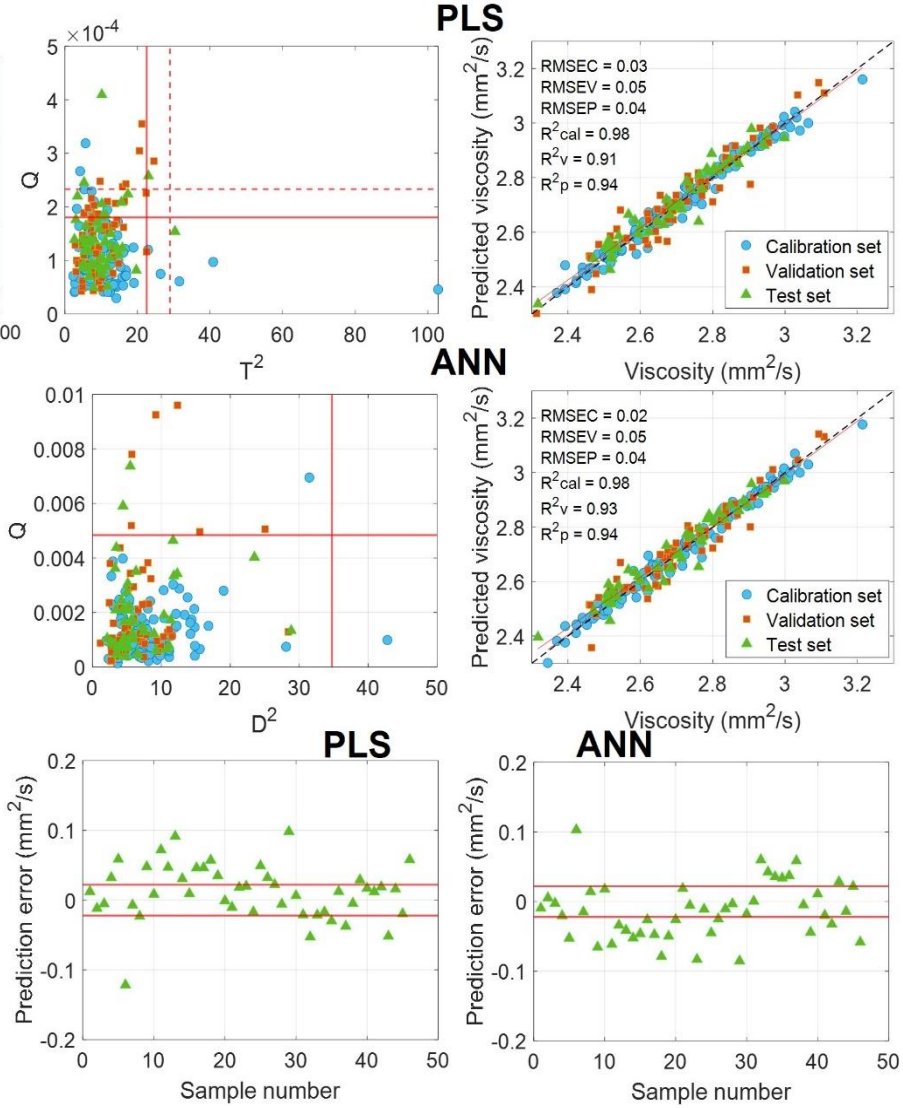


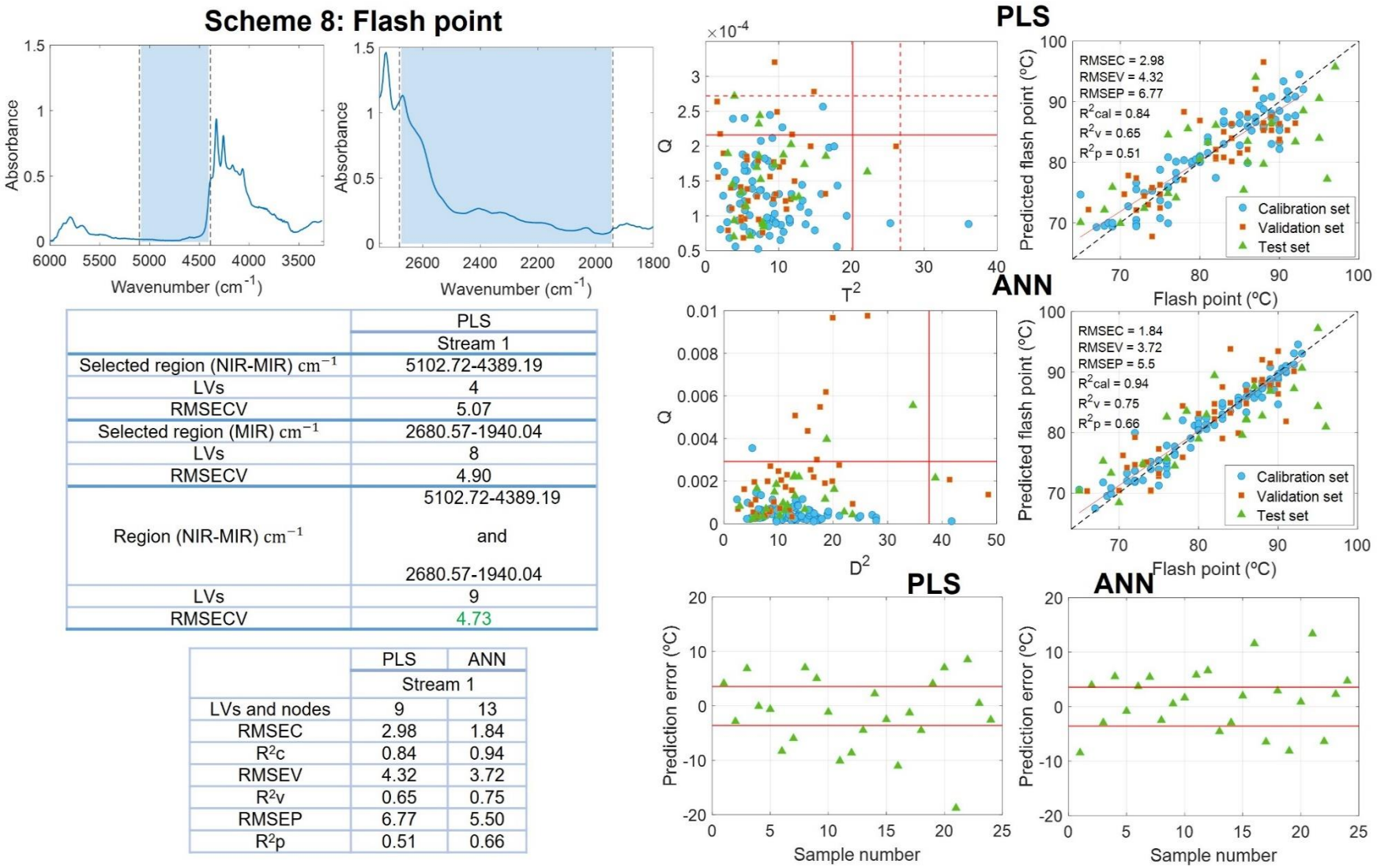
Scheme 7: Viscosity



	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	5461.41-3293.82
LVs	8
RMSECV	0.074
Selected region (MIR) cm^{-1}	2738.42-1820.47
LVs	11
RMSECV	0.038
Region (NIR-MIR) cm^{-1}	5461.41-3293.82 and 2738.42-1820.47
LVs	11
RMSECV	0.042

	PLS	ANN
	Stream 2	
LVs and nodes	11	8
RMSEC	0.03	0.02
R^2c	0.97	0.98
RMSEV	0.05	0.05
R^2v	0.91	0.93
RMSEP	0.04	0.04
R^2p	0.93	0.94

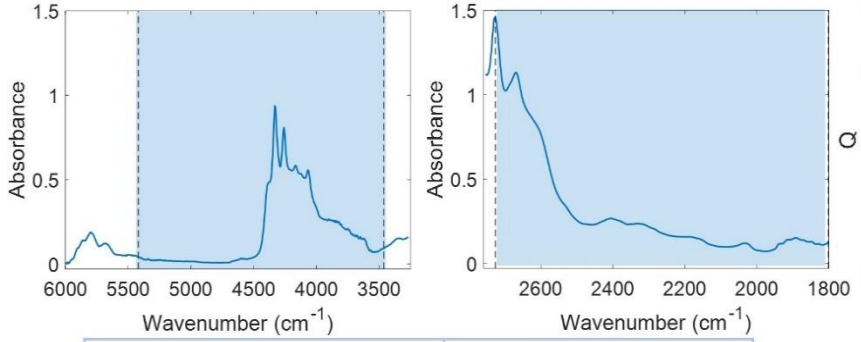


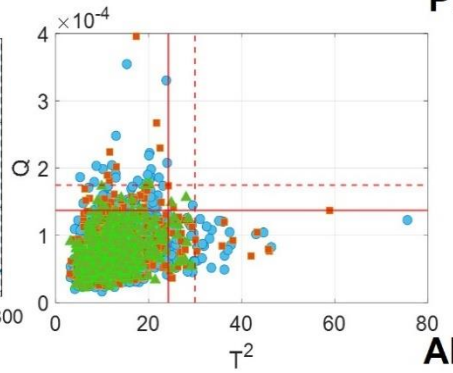


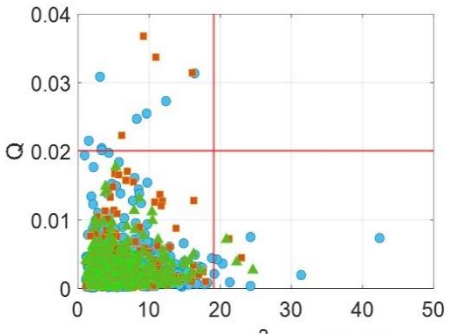
	PLS
	Stream 1
Selected region (NIR-MIR) cm^{-1}	5102.72-4389.19
LVs	4
RMSECV	5.07
Selected region (MIR) cm^{-1}	2680.57-1940.04
LVs	8
RMSECV	4.90
Region (NIR-MIR) cm^{-1}	5102.72-4389.19 and 2680.57-1940.04
LVs	9
RMSECV	4.73

	PLS	ANN
	Stream 1	
LVs and nodes	9	13
RMSEC	2.98	1.84
R^2c	0.84	0.94
RMSEV	4.32	3.72
R^2v	0.65	0.75
RMSEP	6.77	5.50
R^2p	0.51	0.66

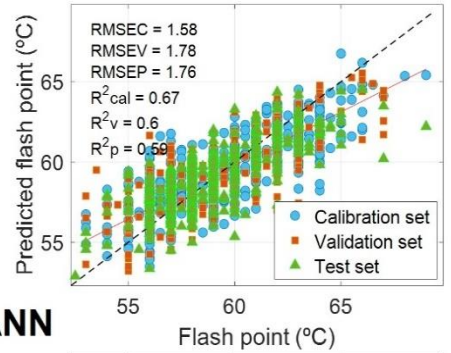
Scheme 9: Flash point



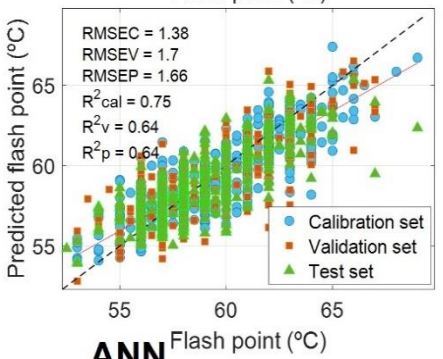




PLS

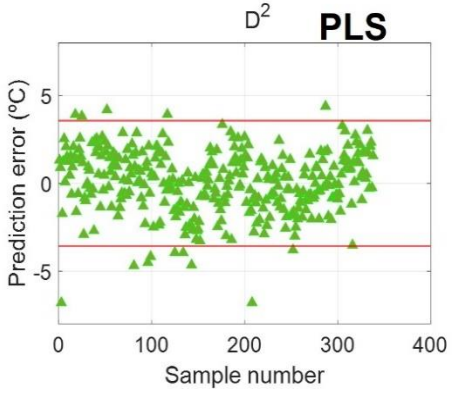


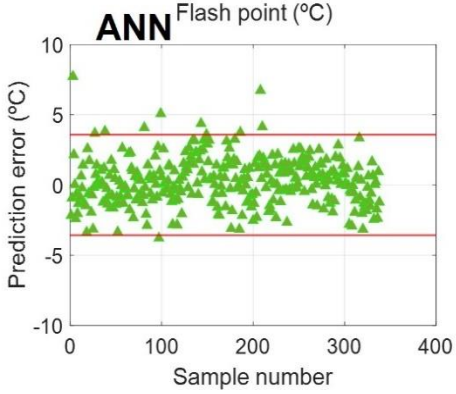
ANN

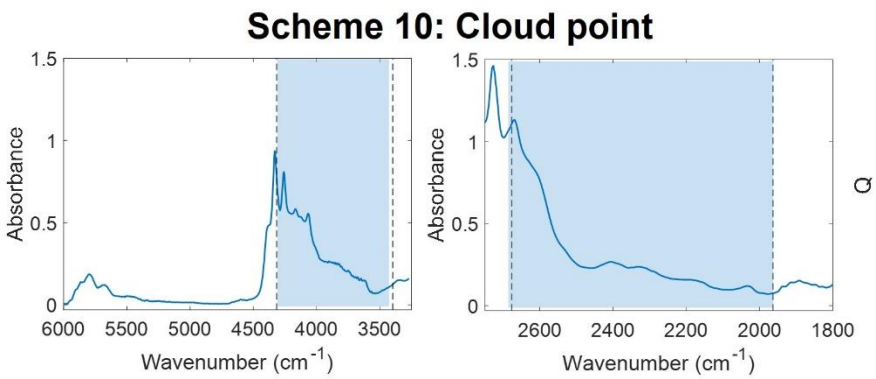


	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	5418.99-3463.53
LVs	9
RMSECV	1.92
Selected region (MIR) cm^{-1}	2726.85-1801.19
LVs	15
RMSECV	1.75
Region (NIR-MIR) cm^{-1}	5418.99-3463.53 and 2726.85-1801.19
LVs	14
RMSECV	1.76

	PLS	ANN
	Stream 2	
LVs and nodes	14	6
RMSEC	1.58	1.38
R^2c	0.67	0.75
RMSEV	1.78	1.70
R^2v	0.60	0.64
RMSEP	1.76	1.66
R^2p	0.59	0.64

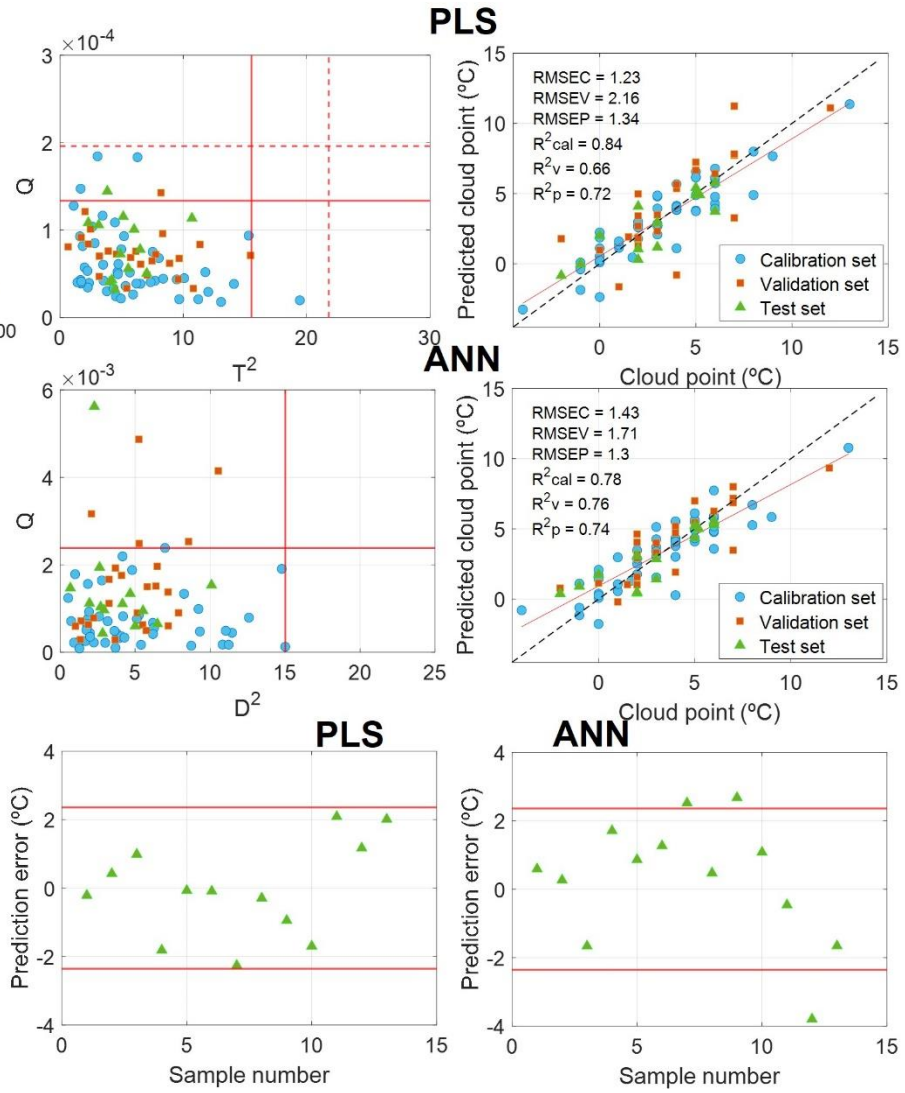




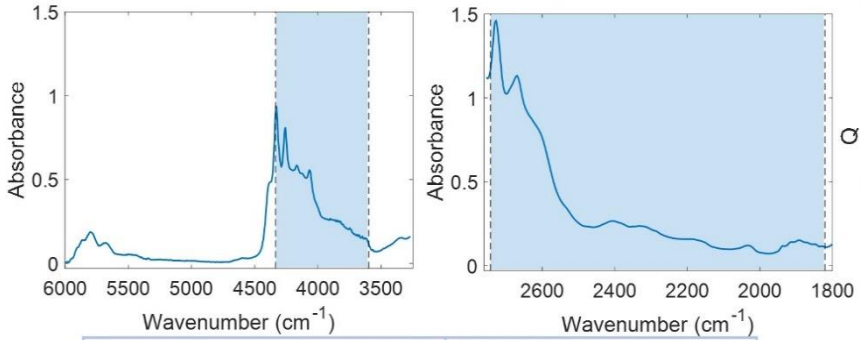


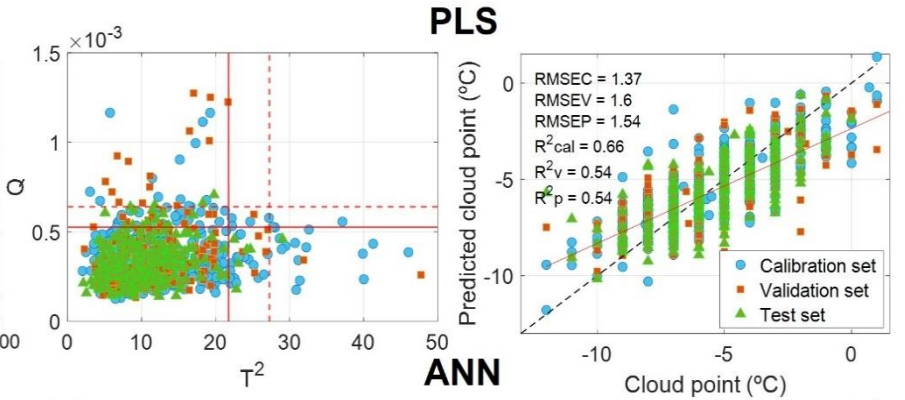
	PLS
	Stream 1
Selected region (NIR-MIR) cm^{-1}	4315.91-3397.96
LVs	6
RMSECV	1.87
Selected region (MIR) cm^{-1}	2676.71-1963.18
LVs	6
RMSECV	1.59
Region (NIR-MIR) cm^{-1}	4315.91-3397.96 and 2676.71-1963.18
LVs	8
RMSECV	1.72

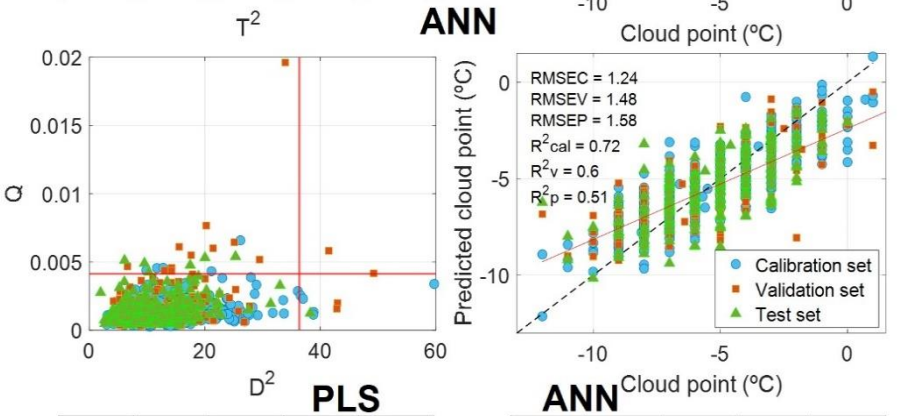
	PLS	ANN
	Stream 1	
LVs and nodes	6	5
RMSEC	1.23	1.43
R^2c	0.84	0.78
RMSEV	2.16	1.71
R^2v	0.66	0.76
RMSEP	1.34	1.30
R^2p	0.72	0.74



Scheme 11: Cloud point

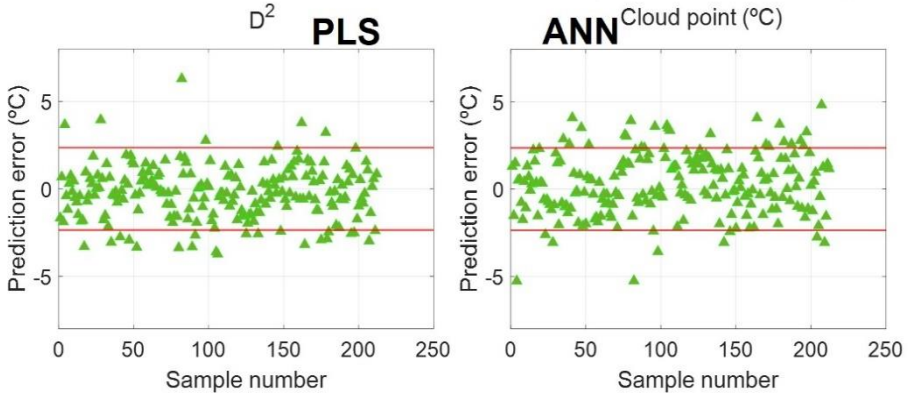


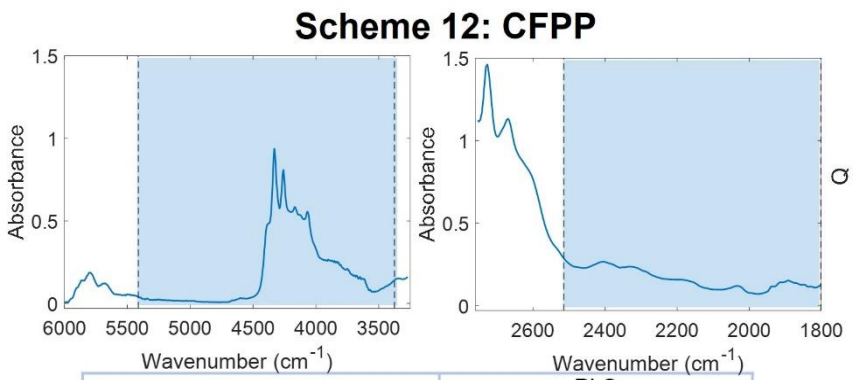




	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4335.19-3598.52
LVs	8
RMSECV	1.63
Selected region (MIR) cm^{-1}	2742.28-1820.47
LVs	17
RMSECV	1.58
Region (NIR-MIR) cm^{-1}	4335.19-3598.52 and 2742.28-1820.47
LVs	12
RMSECV	1.54

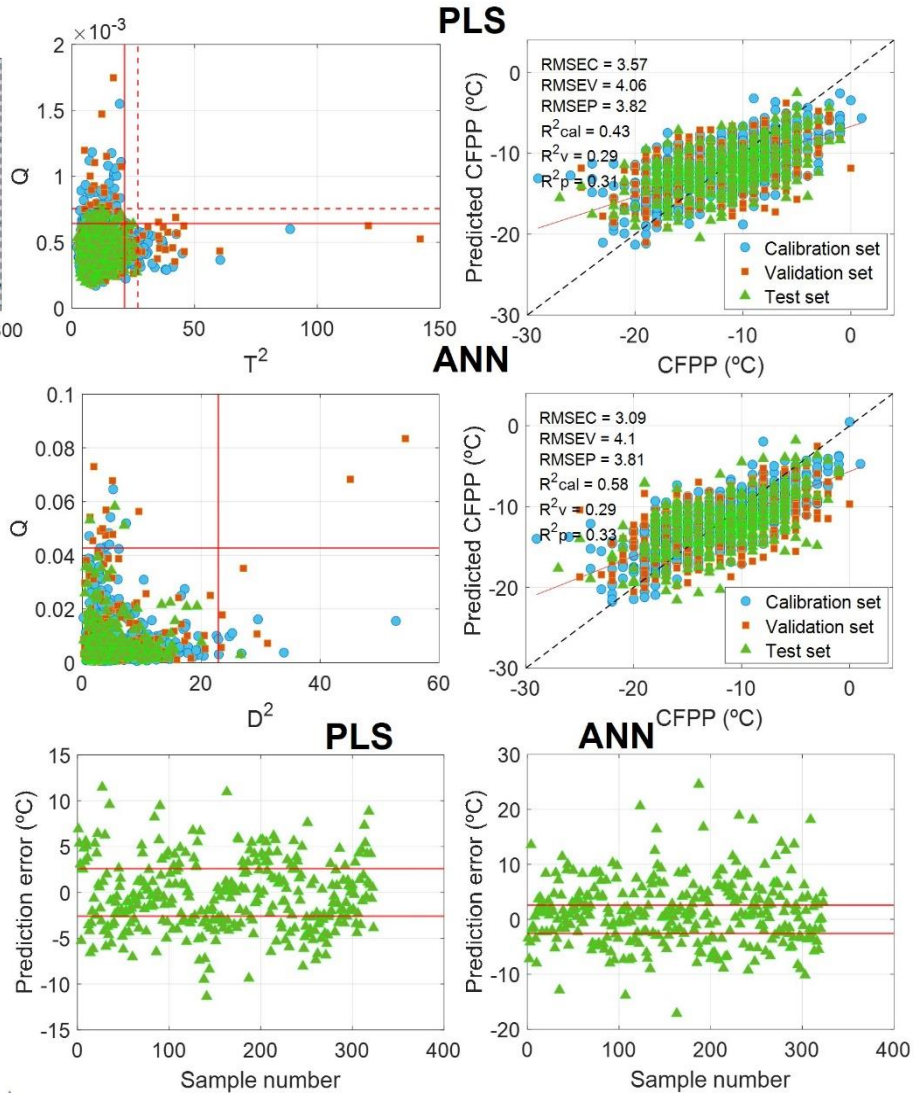
	PLS	ANN
	Stream 2	
LVs and nodes	12	13
RMSEC	1.37	1.24
R^2c	0.66	0.72
RMSEV	1.60	1.48
R^2v	0.54	0.6
RMSEP	1.54	1.58
R^2p	0.54	0.51



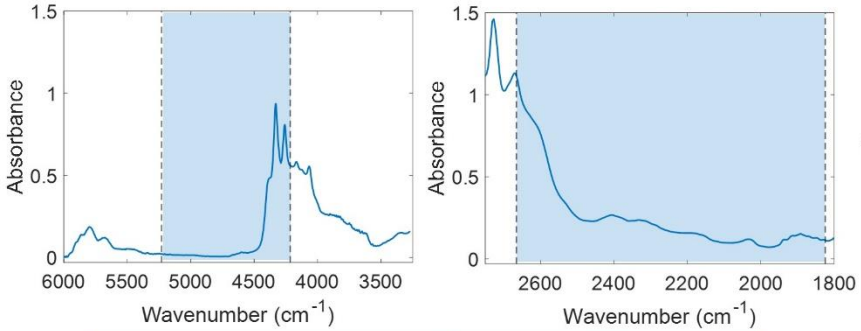


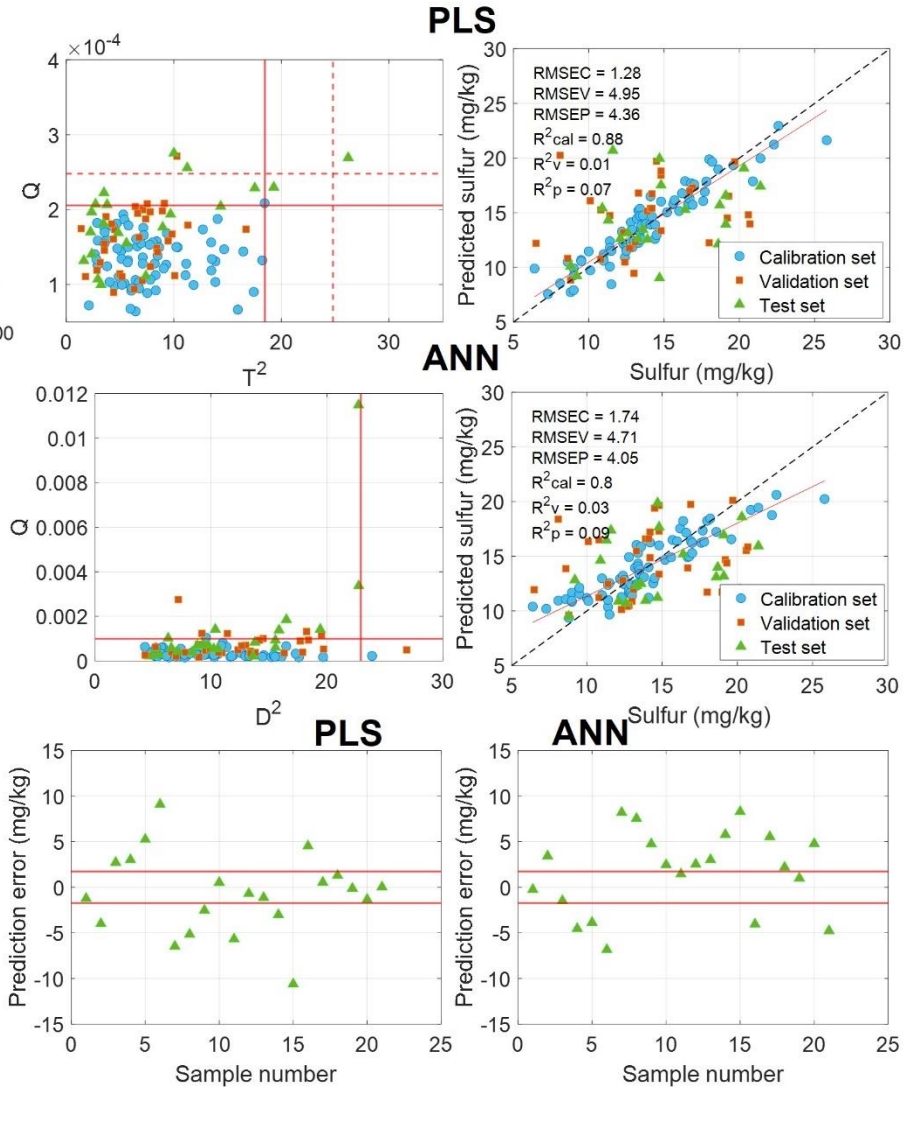
	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	5411.28-3374.82
LVs	9
RMSECV	4.13
Selected region (MIR) cm^{-1}	2514.72-1801.19
LVs	13
RMSECV	4.18
Region (NIR-MIR) cm^{-1}	5411.28-3374.82 and 2514.72-1801.19
LVs	12
RMSECV	4.07

	PLS	ANN
	Stream 2	
LVs and nodes	12	5
RMSEC	3.57	3.09
R^2c	0.43	0.58
RMSEV	4.06	4.10
R^2v	0.29	0.29
RMSEP	3.82	3.81
R^2p	0.31	0.33



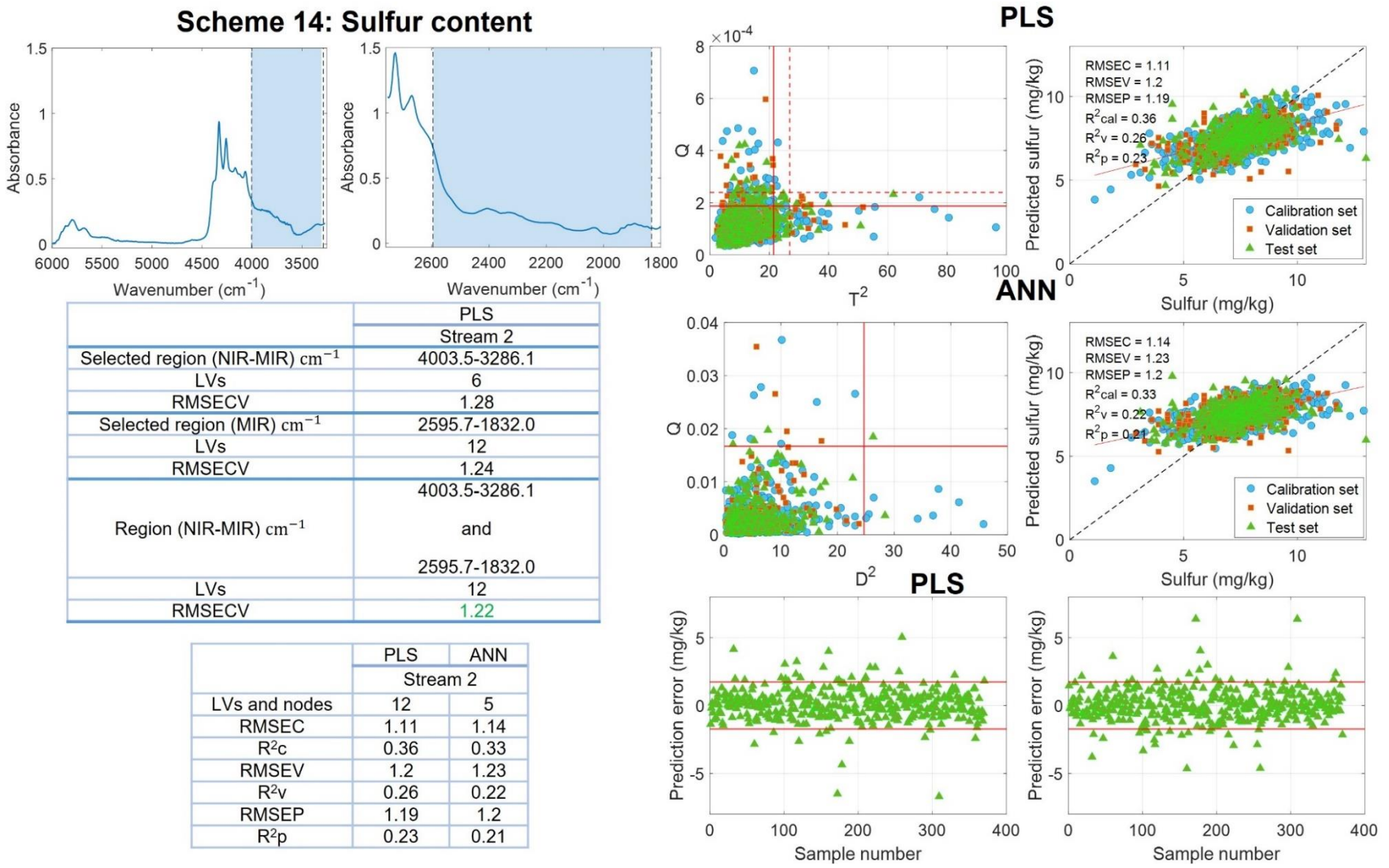
Scheme 13: Sulfur content





	PLS
	Stream 1
Selected region (NIR-MIR) cm^{-1}	5230.0-4215.6
LVs	8
RMSECV	2.45
Selected region (MIR) cm^{-1}	2665.1-1824.3
LVs	15
RMSECV	2.90
Region (NIR-MIR) cm^{-1}	5230.0-4215.6 and 2665.1-1824.3
LVs	11
RMSECV	2.51

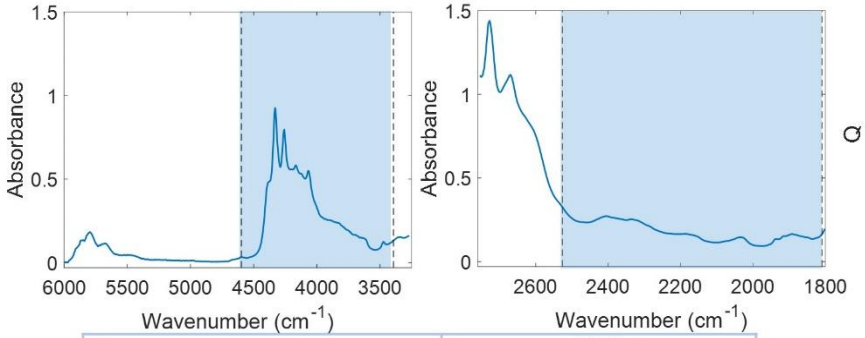
	PLS	ANN
	Stream 1	
LVs and nodes	8	10
RMSEC	1.28	1.74
R^2c	0.88	0.8
RMSEV	4.95	4.71
R^2v	0.01	0.03
RMSEP	4.36	4.05
R^2p	0.07	0.09



	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4003.5-3286.1
LVs	6
RMSECV	1.28
Selected region (MIR) cm^{-1}	2595.7-1832.0
LVs	12
RMSECV	1.24
Region (NIR-MIR) cm^{-1}	4003.5-3286.1 and 2595.7-1832.0
LVs	12
RMSECV	1.22

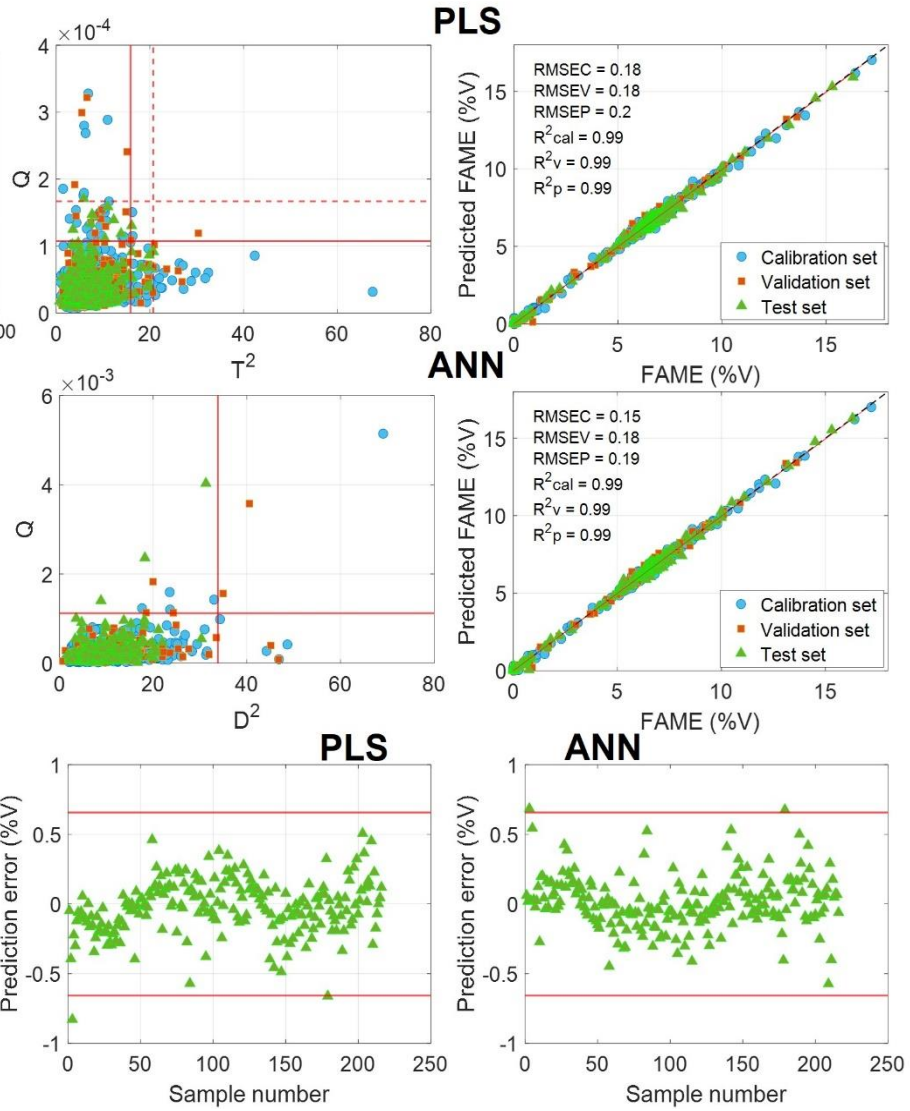
	PLS	ANN
	Stream 2	
LVs and nodes	12	5
RMSEC	1.11	1.14
R^2c	0.36	0.33
RMSEV	1.2	1.23
R^2v	0.26	0.22
RMSEP	1.19	1.2
R^2p	0.23	0.21

Scheme 15: FAME



	PLS
	Stream 2
Selected region (NIR-MIR) cm^{-1}	4597.47-3394.10
LVs	8
RMSECV	0.18
Selected region (MIR) cm^{-1}	2526.29-1808.90
LVs	9
RMSECV	0.17
Region (NIR-MIR) cm^{-1}	4597.47-3394.10 and 2526.29-1808.90
LVs	8
RMSECV	0.17

	PLS	ANN
	Stream 2	
LVs and nodes	8	10
RMSEC	0.18	0.15
R^2c	0.99	0.99
RMSEV	0.18	0.18
R^2v	0.99	0.99
RMSEP	0.20	0.19
R^2p	0.99	0.99



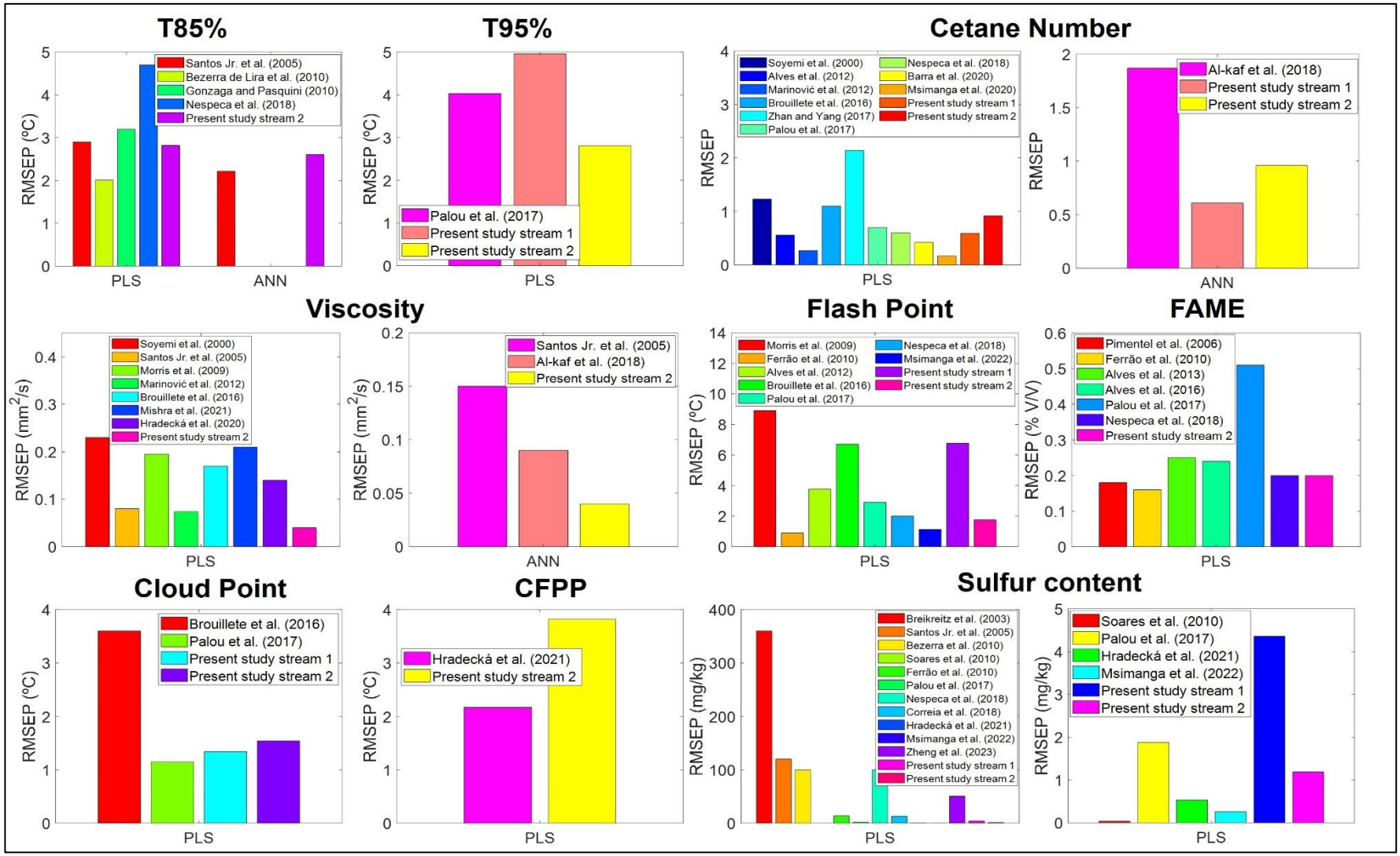


Figure V-17. Comparison of RMSEPs of PLS and ANN for predicting remaining properties with those from the literature.

Partial conclusions

The following conclusions can be drawn from the above PLS and FFNN calibration models:

1. The union of the NIR-MIR and MIR subregions were the optimal spectral regions to develop the PLS model for predicting T65% (4744.03-3579.23 and 2734.57-1801.19 cm^{-1}), T85% (4763.31-3482.81 and 2684.43-1808.90 cm^{-1}), CFPP (5411.28-3374.82 and 2514.72-1801.19 cm^{-1}), flash point of desulfurized samples (5102.72-4389.19 and 2680.57-1940.04 cm^{-1}), cloud point of commercial samples (4335.19-3598.52 and 2742.28-1820.47 cm^{-1}), and sulfur content of commercial samples (4003.5-3286.1 and 2595.7-1832.0 cm^{-1}).
2. The optimal spectral regions to develop the PLS models were the MIR regions between: 2680.57-1855.19 and 2669.0-1801.19 cm^{-1} for predicting T95% recovered of desulfurized and commercial samples, 2618.86-1866.76 cm^{-1} for cetane number of desulfurized samples, 2738.42-1820.47 cm^{-1} for viscosity, 2526.29-1808.90 cm^{-1} for FAME content, 2726.85-1801.19 cm^{-1} for flash point of commercial samples and 2676.71-1963.18 cm^{-1} for cloud point of desulfurized samples, respectively.
3. Only for predicting the cetane number of commercial samples and sulfur content of desulfurized samples, the optimal spectral regions were the NIR-MIR region between 4400.76-3270.68 cm^{-1} and 5230-4215.6 cm^{-1}, respectively.
4. The RMSEPs values of the PLS models for predicting T65%, T85%, T95%, flash point, cloud point, density, cetane number, sulfur content, viscosity, CFPP, and FAME of commercial samples were 2.83 ºC, 2.82 ºC, 2.81ºC, 1.76 ºC, 1.30 ºC, 0.46 kg/m^3, 0.92, 1.19 mg/kg, 0.04 mm^2/s, 3.82 ºC, and 0.20 % (v/v), respectively. The corresponding values of FFNN model were 2.75 ºC, 2.61 ºC, 2.92ºC, 1.66 ºC, 1.58 ºC, 0.42 kg/m^3, 0.96, 1.20 mg/kg, 0.04 mm^2/s, 3.81 ºC, and 0.19 % (v/v).
5. The RMSEPs values of the PLS models for predicting T95%, flash point, cloud point, density, cetane number, and sulfur content of desulfurized samples were 4.96ºC, 6.77 ºC, 1.34 ºC, 0.73 kg/m^3, 0.59 and 4.36 mg/kg, respectively. The corresponding values of FFNN models were 4.65ºC, 5.50 ºC, 1.30 ºC, 0.85 kg/m^3, 0.61 and 4.05 mg/kg, respectively.
6. To the best of the author's knowledge, our models achieved the lowest RMSEP for viscosity. For the other properties, the obtained RMSEP values are generally satisfactory and comparable with the best results in the literature.

V.3.3 PLS vs ANNs: general remarks

Figure V-18 summarises the results obtained by PLS and ANN for the prediction of diesel properties. The ANN models demonstrated a superior predictive ability compared to the PLS models for the following four groups of properties: **1)** T65%, T85%; **2)** T95%, cloud point, and sulfur content of samples from stream 1; **3)** the density of samples from stream 2; and **4)** the flash point of samples from each stream. Conversely, the PLS models demonstrated superior predictive ability over ANN models for **1)** the density of samples from stream 1 and **2)** T95% and cloud point of stream 2. Furthermore, the predictive ability of both PLS and ANN models was very similar for the cetane number of samples from streams 1 and 2, viscosity, CFPP, the sulfur content of samples from stream 2, and FAME content.

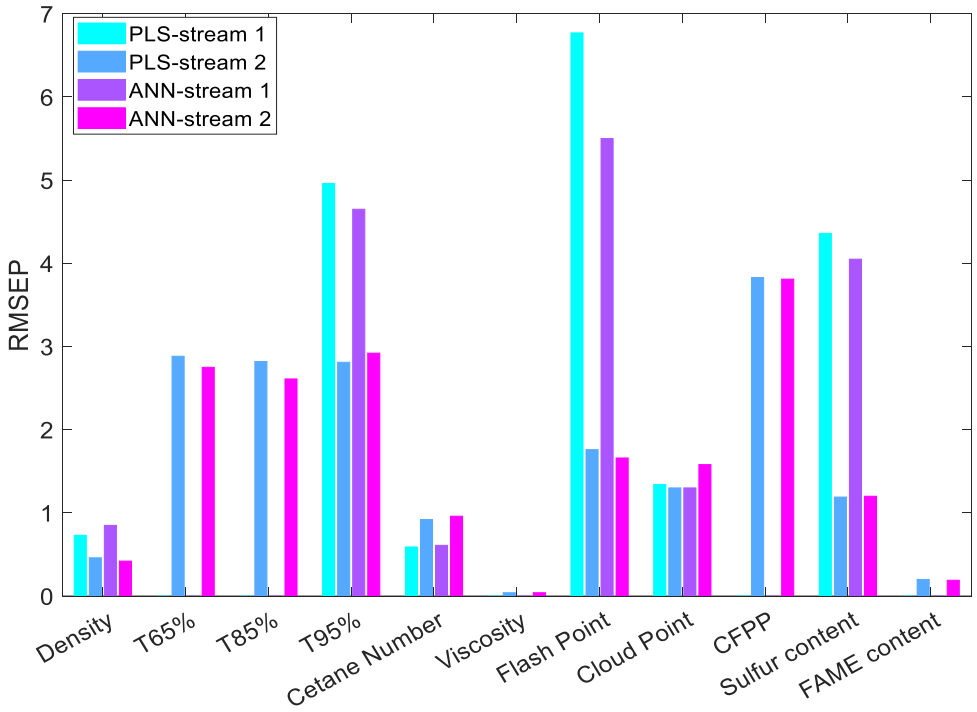


Figure V-18. Comparison of the performance of PLS and ANN calibration model in terms of root mean squared errors of prediction (RMSEP) values for eleven diesel properties: density (at 15 ºC), distillation temperatures at 65%, 85%, and 95%, cetane number, kinematic viscosity (at 40 ºC), flash point, cloud point, CFPP, sulfur content, and FAME content.

The results showed that, despite the potential of ANN models to capture complex nonlinear relationships in data, making them suitable for tackling intricate patterns anticipated challenging for linear methods, PLS models provided comparable results for most properties.

It should also be noted that, in the present study, the predictive ability appears to rely more on the type of the samples (stream 1 or stream 2) and, thus, on the nature of the relationship between the IR spectrum and property than on the chosen approach for modeling this relationship. In this sense, except for the cetane number, the predictive abilities of both PLS and ANN models were superior for samples from stream 2. This behavior is noticeable for density, T95%, flash point, and sulfur content. A possible explanation for this behavior could be that the models are consistently developed with a more significant number of samples from stream 2 than from stream 1.

Regarding the cetane number, it is plausible that the samples from stream 2, being a finished product, may include specific additives such as a cetane improver. Cetane improver is utilized to raise the cetane number, thereby improving the ignition properties of diesel fuel. Since detailed information on the possible addition of the cetane improver was not available, its chemical composition and concentration in the samples of stream 2 were unknown. Consequently, the models developed did not consider the possible spectral variability associated with this additive, which could have affected the predictive ability for cetane number in the samples of stream 2.

Concerning parameters such as interpretability, sample efficiency, sensitivity to outliers, ease of use, computational efficiency, and stability, both modeling approaches can be organized in the following order:

Interpretability: PLS > ANN. The interpretability of PLS models was often more straightforward and intuitive than that of ANN models. This is probably attributed to the PLS model directly identifying LVs with meaningful physical interpretations, thereby providing insights into the IR spectrum-property relationship.

Sample efficiency: PLS > ANN. PLS models exhibit satisfactory performance even when trained on a limited number of samples, while ANN models require larger datasets to achieve successful training.

Sensitivity to detecting outliers: ANN > PLS. ANN models were more sensitive to detecting outliers in the data than PLS models.

Ease of use: PLS > ANN. PLS models are characterized by their simplicity and require less tuning compared to ANN models. ANN required more model hyperparameters to optimize.

Computational efficiency: PLS > ANN. ANN training required more computational (CPU) times compared to the model-building approaches of PLS.

Stability: PLS > ANN. PLS models exhibit better stability to minor variations in spectral data, while the ANN model may be more susceptible to overfitting in such scenarios.

Table V-3 summarizes the percent of test samples that fall in the range established by Eq. (II.13) using the reference and estimated values by the PLS and ANN models. Only models with R^2 >0.75 were considered. One can notice that in addition to the IR/PLS and IR/ANN models for the density of commercial samples, only the PLS and ANN models for FAME content can be considered valid for practical implementation in routine diesel analysis. Regarding FAME content, merely 0.9% of the test samples for both the PLS and ANN models exhibited prediction errors exceeding the tolerance limits established by the reference method.

Table V-3. Percentage of the test samples that fall within the tolerance limits admitted for each property.

Property	PLS		ANN	
	Stream 1	Stream 2	Stream 1	Stream 2
Density	79.2	95.6	75	97.1
T65%	-	68.3	-	71
T85%	-	[a]	-	[a]
T95%	[a]	[a]	[a]	[a]
Cetane Number	90	[a]	90	[a]
Viscosity	-	50	-	41.3
Flash Point	[a]	[a]	41.7	[a]
Cloud Point	[a]	86.8	[a]	[a]
CFPP	-	[a]	-	[a]
Sulfur content	[a]	[a]	[a]	[a]
FAME content	-	99.1	-	99.1

[a] Models with R^2-values < 0.75 were not considered.

It is important to emphasize that while the models have a high predictive ability for cetane number and viscosity, the admitted tolerance limits are very narrow, making it susceptible to exceeding them. Conversely, although the predictive ability is lower for properties such as cloud point, the broader tolerance limits result in a high percentage of samples within the acceptable range.

Overall, it is essential to recognize that ASTM reference analyses incorporate diverse measurement errors, which limit the predictive ability of the developed chemometric models. In this regard, the usefulness of the developed or any predictive models should be evaluated considering the expected predictive ability for the given application for which they are intended. Thus, predictive models exhibiting high

prediction error compared to the ASTM reference method might not serve as a replacement for that method. Still, instead, they could prove highly valuable as a detection tool to flag potentially suspicious samples for further analysis [17].

V.4 Conclusions

The following conclusions can be drawn from the development of both PLS and ANN calibration models for each property:

1. PLS and FFNN models employing IR spectra yielded accurate predictions for density, cetane number, viscosity, and FAME content. Both models for distillation temperatures, flash point, cloud point, CFPP, and sulfur content were not satisfactory enough.
2. The ranges of the properties corresponding to the calibration models for commercial samples were 820-850 kg/m^3 (density), 265-320 ºC (T65%), 305-350 ºC (T85%), 340-370 ºC (T95%), 48-57 (cetane number), 2.3-3.3 mm^2/s (viscosity), 52-70 ºC (flash point), -12.5-2.5 ºC (cloud point), -30-2.5ºC (CFPP), 0-13 mg/kg (sulfur content) and 0-17.5 %v/v (FAME content). For desulfurized samples, the corresponding ranges of the properties were 835-875 kg/m^3 (density), 350-390 ºC (T95%), 50-60 (cetane number), 60-100 ºC(flash point), -5-15 ºC (cloud point), 5-30 mg/kg (sulfur content).
3. Superior predictive performance of the ANN model over the PLS model was found for T65%, T85%; T95%, cloud point, and sulfur content of samples from stream 1; the density of samples from stream 2; and the flash point of samples from each stream. In contrast, the PLS model outperforms the ANN model for the density of samples from stream 1 and T95% and cloud point of samples from stream 2. The predictive ability of ANN and PLS models was similar for the remaining properties. Each method had its strengths and weaknesses, and the most suitable method should be selected based on the specific characteristics of the dataset, the complexity of the relationship between variables, and the requirements of the application of interest.

5. Based on ASTM-E-1655, the IR/PLS and IR/ANN models can be used for determining density and FAME content in samples from stream 2 for practical implementation in routine analysis.

V.5 References

[1] L. de F. Bezerra de Lira, F.V. Cruz de Vasconcelos, C. Fernandes Pereira, A.P. Silveira Paim, L. Stragevitch, M.F. Pimentel, Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration, Fuel. 89 (2010) 405–409. https://doi.org/10.1016/j.fuel.2009.05.028.

[2] M.F. Ferrão, M.D.S. Viera, R.E.P. Pazos, D. Fachini, A.E. Gerbase, L. Marder, Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions, Fuel. 90 (2011) 701–706. https://doi.org/10.1016/j.fuel.2010.09.016.

[3] S. Marinović, M. Krištović, B. Špehar, V. Rukavina, A. Jukić, Prediction of diesel fuel properties by vibrational spectroscopy using multivariate analysis, J. Anal. Chem. 67 (2012) 939–949. https://doi.org/10.1134/S1061934812120039.

[4] M.G. Nespeca, R.R. Hatanaka, D.L. Flumignan, E. De Oliveira, Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis, J. Anal. Methods Chem. (2018) 1795624. https://doi.org/10.1155/2018/1795624.

[5] V.O. Santos Jr., 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. 547 (2005) 188–196. https://doi.org/10.1016/j.aca.2005.05.042.

[6] H.A.G. Al-kaf, K.S. Chia, N.A.M. Alduais, A comparison between single layer and multilayer artificial neural networks in predicting diesel fuel properties using near infrared spectrum, Pet. Sci. Technol. 36 (2018) 411–418. https://doi.org/10.1080/10916466.2018.1425717.

[7] F.B. Gonzaga, C. Pasquini, A low cost short wave near infrared spectrophotometer: Application for determination of quality parameters of diesel fuel, Anal. Chim. Acta. 670 (2010) 92–97. https://doi.org/10.1016/j.aca.2010.04.060.

[8] A. Palou, A. Miró, M. Blanco, R. Larraz, J.F. Gómez, T. Martínez, J.M. González, M. Alcalà, Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 180 (2017) 119–126. https://doi.org/10.1016/j.saa.2017.03.008.

[9] J.C.L. Alves, C.B. Henriques, R.J. Poppi, Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system, Fuel. 97 (2012) 710–717. https://doi.org/10.1016/j.fuel.2012.03.016.

[10] C. Brouillette, W. Smith, C. Shende, Z. Gladding, S. Farquharson, R.E. Morris, J.A. Cramer, J. Schmitigal, Analysis of Twenty-Two Performance Properties of Diesel, Gasoline, and Jet Fuels Using a Field-Portable Near-Infrared (NIR) Analyzer, Appl. Spectrosc. 70 (2016) 746–755. https://doi.org/10.1177/0003702816638279.

[11] O.O. Soyemi, M.A. Busch, K.W. Busch, Multivariate Analysis of Near-Infrared Spectra Using the G-Programming Language, J. Chem. Inf. Comput. Sci. 40 (2000) 1093–1100. https://doi.org/10.1021/ci000447r.

[12] R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, S.L. Rose-Pehrsson, Rapid fuel quality surveillance through chemometric modeling of near-infrared spectra, Energy and Fuels. 23 (2009) 1610–1618. https://doi.org/10.1021/ef800869t.

[13] P. Mishra, F. Marini, A. Biancolillo, J.M. Roger, Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques, Talanta. 223 (2021) 121693. https://doi.org/10.1016/j.talanta.2020.121693.

[14] I. Hradecká, R. Velvarská, K.D. Jaklová, A. Vráblík, Rapid determination of diesel fuel properties by near-infrared spectroscopy, Infrared Phys. Technol. 119 (2021) 103933. https://doi.org/10.1016/j.infrared.2021.103933.

[15] I. Pinheiro Soares, T. F. Rezende, I.C. P. Fortes, Determination of sulfur in diesel using ATR/ FTIR and multivariate calibration, Eclet. Quim. 35 (2010) 71–78. https://doi.org/10.26850/1678-4618EQJ.V35.2.2010.P71-78.

[16] H.Z. Msimanga, C.R. Dockery, D.D. Vandenbos, Classification of local diesel fuels and simultaneous prediction of their physicochemical parameters using FTIR-ATR data and chemometrics, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 279 (2022) 121451. https://doi.org/10.1016/j.saa.2022.121451.

[17] K.J. Johnson, J. Schmitigal, G.J. Walker, NIR Spectrometry and Fuel Quality Assessment, Washington, United States, 2019. https://apps.dtic.mil/sti/pdfs/AD1089383.pdf.

Chapter VI
Monitoring and maintenance of PLS models[2]

[2] Section VI.3 is adapted from M. S. Rodríguez Barrios, E. Ruiz, M.S. Larrechi, J. Ferré, *Comparative analysis of the performance of different approaches for the adaptation of a calibration model for diesel analysis*, Infrared Phys. Technol. (Accepted with moderate and minor revisons).

Chapter VI

VI.1 Introduction

As mentioned in Chapter II, calibration models must be monitored over time to ensure that they maintain their predictive ability. Degradation in performance is typically caused by the so-called instrumental drift and product drift [1]. In diesel analysis, monitoring the predictive ability of the calibration model over time is crucial since products, production processes, or measuring instruments may not remain stable in time. For example: 1) diesel composition may vary over time due to changes in feedstock (i.e., crude oil), refining processes, or environmental regulations; 2) the spectrophotometer used for the calibrations may be renewed, or its quality may deteriorate; 3) external factors, such as temperature and humidity, can influence the properties of diesel samples. After identifying the new sources of variation in the data, strategies for adapting the model can be selected. These may include spectral correction, model parameter tunning, or model expansion [2].

This chapter is divided into two main sections. In Section VI.2, a strategy for monitoring the performance over time of the PLS models developed in Chapter V is presented. In section VI.3, three approaches to correct instrumental drifts between two spectrophotometers are considered and compared.

VI.2 Monitoring the predictive ability of PLS calibration models over time

New characteristics of incoming samples could be identified by comparing model diagnostic measures, such as Hotelling's T^2 and Q residuals, to the established limits of applicability for a calibration model. Samples with high T^2 values often exhibit a chemical composition similar to the calibration samples but with extreme concentrations or unusual combinations thereof [3]. Samples with high Q residual represent singular samples with new analytes or other new variations (such as changes in the measurement conditions or the instrument).

Below, the methodology adopted for monitoring the predictive ability of the PLS calibration models over time is addressed. This methodology was based on Hotelling's T^2 and Q-residuals and the prediction errors. Prediction errors outside the tolerance limits admitted by the reference method or biased predictions indicate degradation of the model's predictive ability over time due to product drift. Moreover, it was assumed that there was no instrumental drift over the time period considered in the analysis performed in this section.

Materials and methods

The incoming samples set consisted of 571 samples produced between 2021 and 2022. The corresponding spectra were measured using the procedure described in section III.4.1 of Chapter III. Only a third of these samples were used to detect deviations in the predictions of the calibration model over time. The samples were selected by taking the first of every three samples in the order in which they were produced. The order in which the samples were produced is retained to examine the impact of emergent variations on prediction performance. Only these diesel samples were analyzed using the reference methods listed in section I.3 of Chapter I.

The adopted methodology consisted of two main steps:

1) Monitoring Hotelling's T^2 and Q-residuals of incoming samples considering the limits of applicability of PLS/IR models developed in section V.2.4 of Chapter V.
2) To verify that the prediction errors remain constant and uniformly distributed over time. The prediction error of each sample was compared to the tolerance limits admitted by the reference method for each property as described in section V.2.3 in Chapter V. More than 5% of samples with prediction errors beyond the tolerance limits admitted by the reference method were considered as an indicator of the invalidity of the model for routine diesel analysis.

Results and discussion

Figure VI-1 shows the results of monitoring the PLS model of density applied to incoming samples from streams 1 (left) and 2 (right), expressed in terms of Hotelling's T^2 and Q-residuals. It can be seen that some incoming samples for both streams exceeded one of the two limits. Yet, for a significance level of 99%, no sample exceeded Q_{lim} and T^2_{lim} simultaneously. This could suggest that the variability of these incoming samples was included in the calibration model.

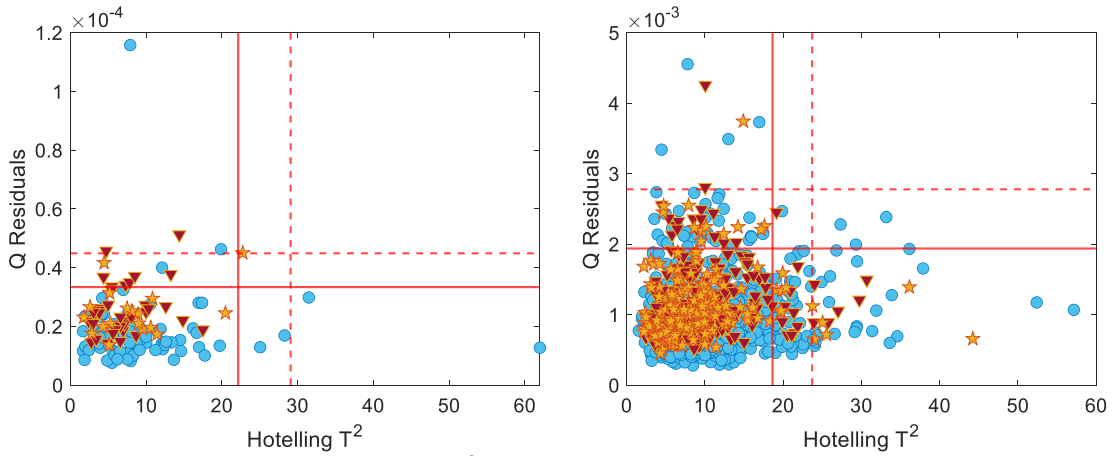


Figure VI-1. Q Residuals vs. Hotelling T^2 for the PLS model of density applied to samples from stream 1 (left) and stream 2 (right). Solid and dashed lines represent the limit of Hotelling's T^2 and Q statistics at 95% and 99% confidence. Calibration samples (blue circles), incoming samples analyzed using the reference methods (brown inverted triangles), and remaining incoming samples (golden star).

Figure VI-2 shows the prediction errors of density of the selected incoming samples from streams 1 and 2 in temporal order over the entire control period (10 months) and the allowed tolerance limit. It is observed that there is no clear temporal trend in the prediction error. The RMSEP (1.04 kg/m^3 and 0.64 kg/m^3) and average residual values (0.53 kg/m^3 and -0.09 kg/m^3) for each stream were not markedly high, being higher for stream 1.

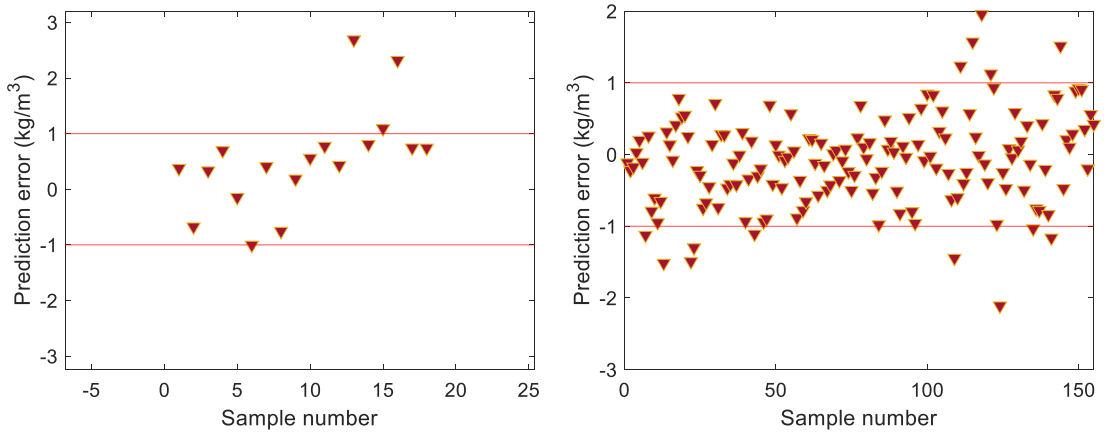
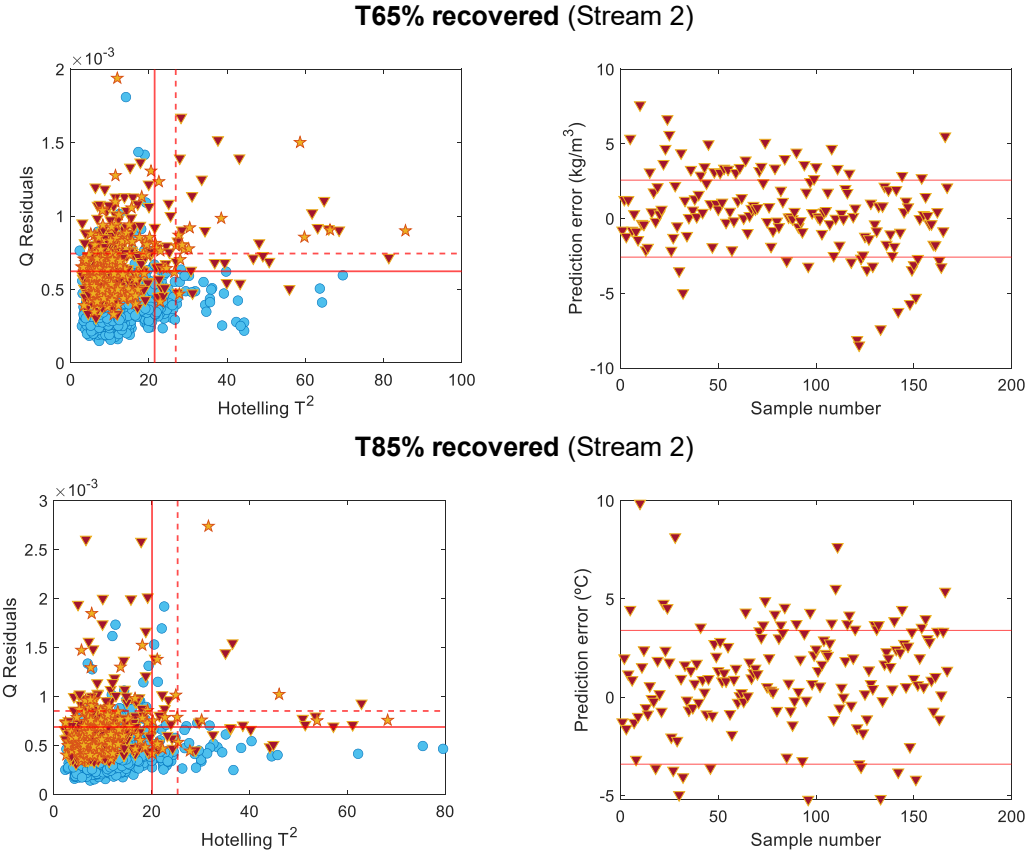


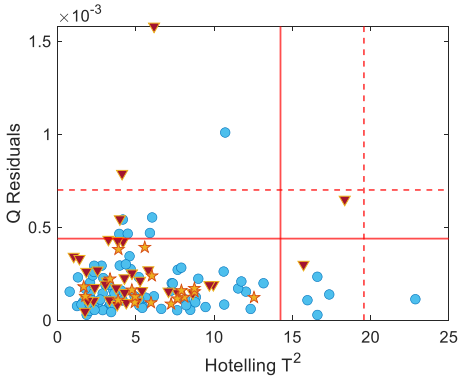
Figure VI-2. Density prediction error of the incoming samples from streams 1 (left) and 2 (right) in temporal order.

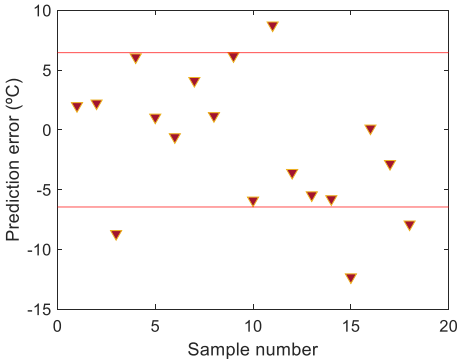
Overall, of the selected incoming set, only 3 (~17%) and 14 (~9%) of the samples from streams 1 and 2, respectively, were outside the admitted tolerance limits. These results might suggest that the model retained, to some extent, its stability over time. However, based on ASTM-E-1655, none of the density models remained valid for routine analysis after ten months due to the percentage of samples that exceeded the limits admitted by the reference method. To deal with this issue, updating the models can be considered.

The results of monitoring the stability of PLS models for the remaining properties are shown below in Figure VI-3. It is possible to observe that the variability of incoming samples was included in the calibration model for the cetane number and cloud point of samples from streams 2 and 1, respectively. In contrast, some random samples over time were representative of new sources of variability in the production process (novel spectra) not included in the calibration model to predict the remaining properties. The novel spectra were not the same for all properties. Also, for flash point, FAME content, and sulfur content, a cluster of incoming samples with high T^2 and Q beyond the limits of applicability was observed. Except for flash point (stream 2), sulfur content (stream 1), and FAME content, monitoring prediction errors revealed no temporal trend.

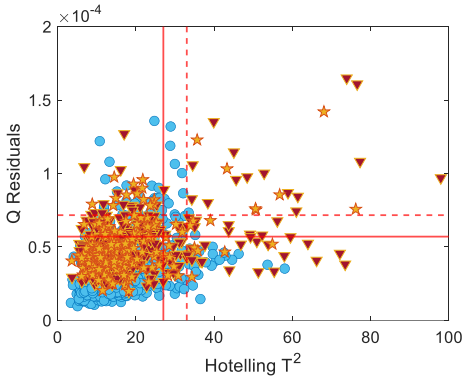


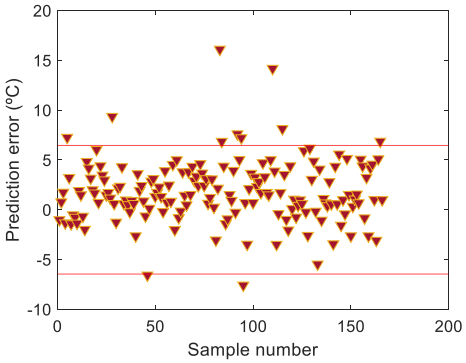
T95% recovered (Stream 1)



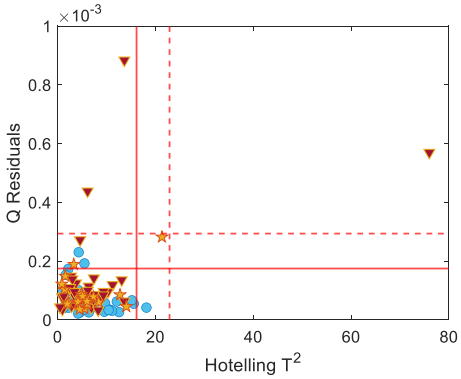


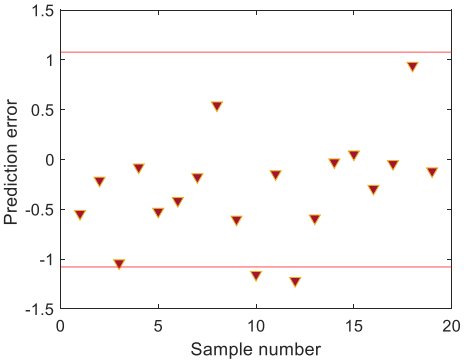
T95% recovered (Stream 2)



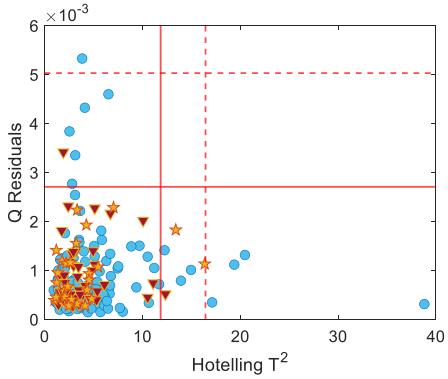


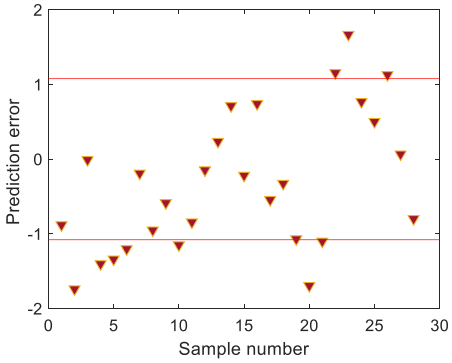
Cetane Number (Stream 1)



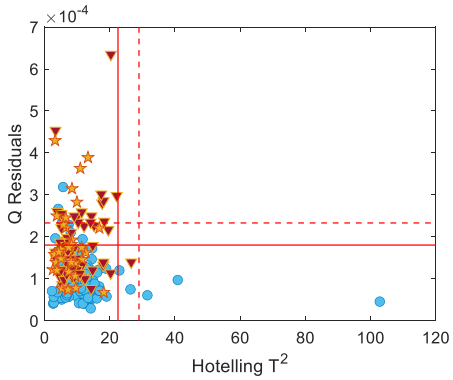


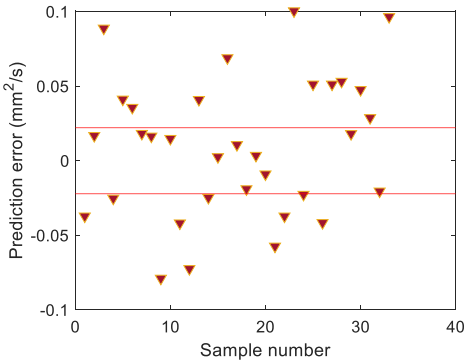
Cetane Number (Stream 2)



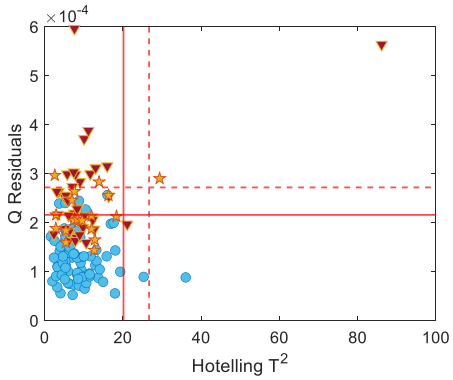


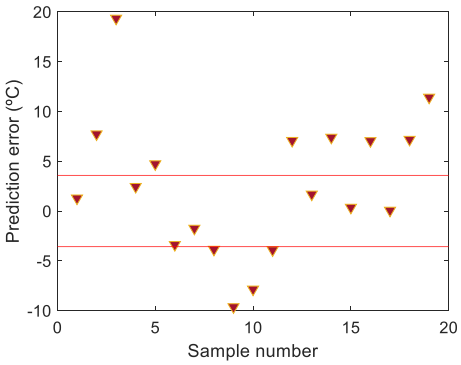
Viscosity (Stream 2)



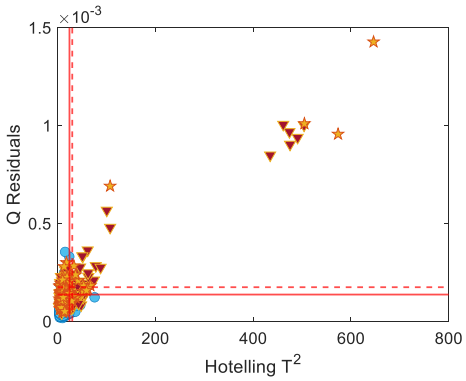


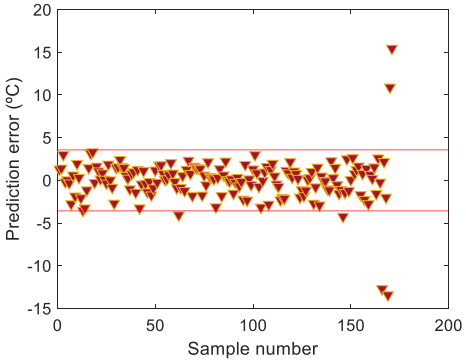
Flash point (Stream 1)



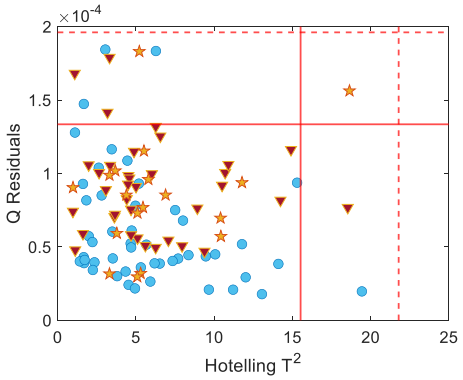


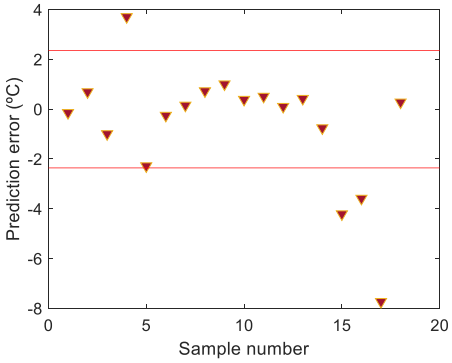
Flash point Stream 2



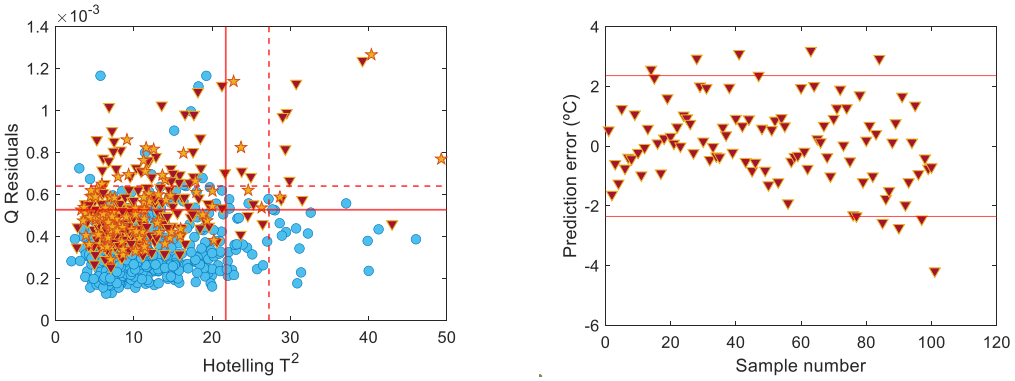


Cloud point (Stream 1)

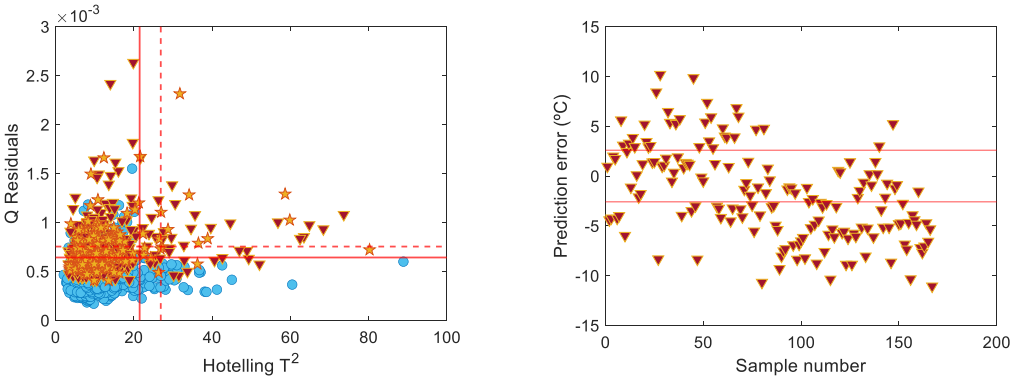




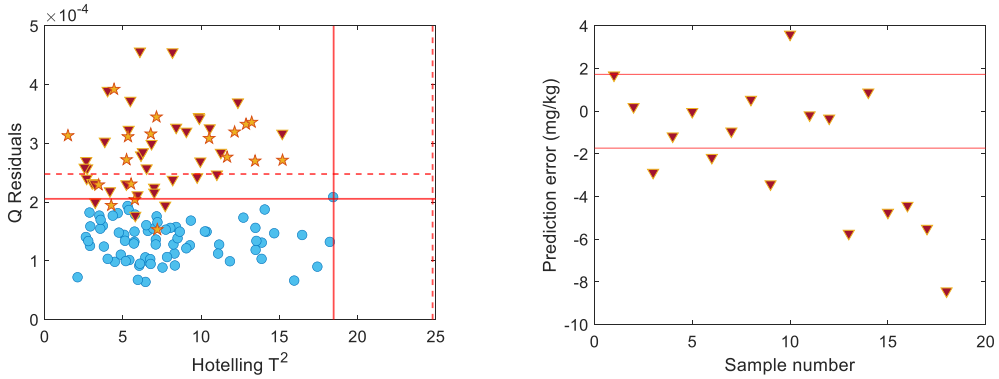
Cloud point (Stream 2)



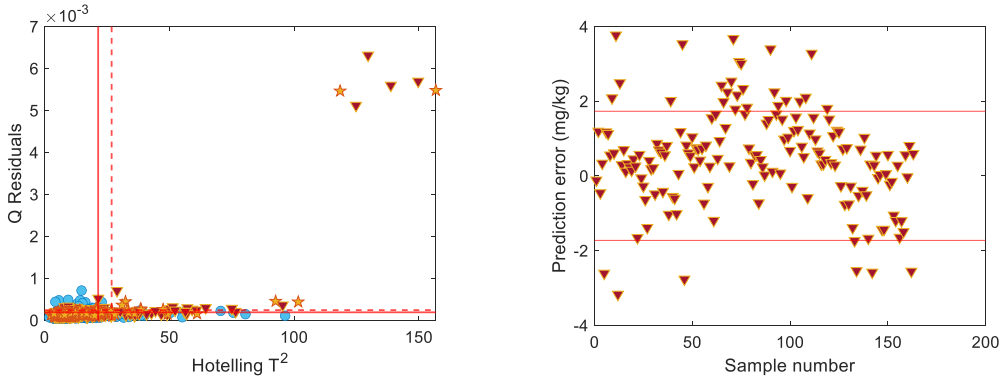
CFPP (Stream 2)



Sulfur content (Stream 1)



Sulfur content (Stream 2)



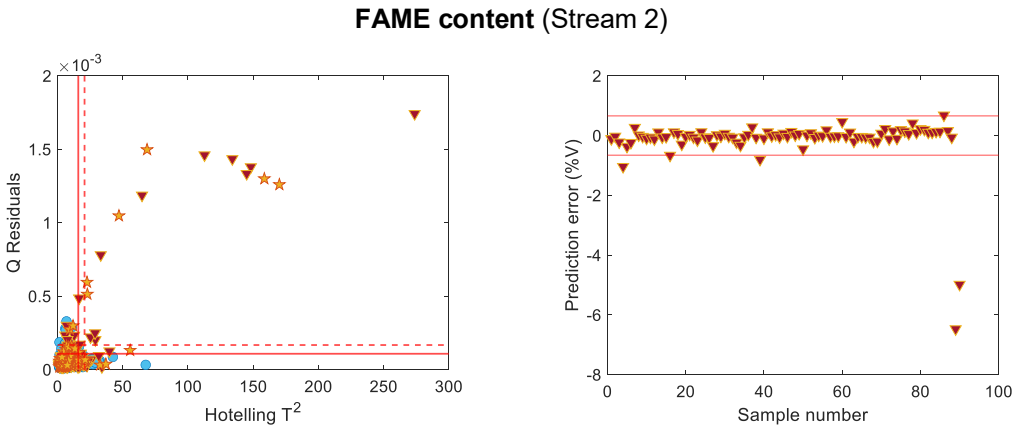


Figure VI-3. Results of stability control of PLS models of the remaining properties.

Table VI-1 lists the values of average residuals and RMSEP for all properties. It can be observed that the largest values corresponds to T95%, flash point from stream 1, and CFPP. Based on the referred ASTM, only the model for predicting FAME content retained its validity for routine analysis after 10 months, since 96% of the new samples fall within the admitted tolerance limits (Table VI-2). For this reason, although the calibration model withstands new variations of incoming samples, their maintenance must be also considered.

Table VI-1. Residual statistics of predictions of the PLS model for samples collected over 10 months.

Property	Average residuals		RMSEP	
	Stream 1	Stream 2	Stream 1	Stream 2
Density (kg/m^3)	0.53	-0.09	1.04	0.64
T65 % (ºC)	-	0.38	-	2.66
T85 % (ºC)	-	1.07	-	2.70
T95 % (ºC)	-1.19	1.89	5.72	3.59
Cetane Number	-0.30	-0.33	0.60	0.96
Viscosity (mm^2/s)	-	0.0096	-	0.0470
Flash Point (ºC)	2.49	0.03	7.26	2.56
Cloud Point (ºC)	-0.67	0.11	2.51	1.36
CFPP (ºC)	-	-1.83	-	4.90
Sulfur content (mg/kg)	-1.83	0.47	3.50	1.36
FAME content (%v/v)	-	-0.15	-	0.89

Table VI-2. Percentage of incoming samples that fall within the tolerance limits admitted for each property during the control of model stability.

Diesel property	PLS	
	Stream 1	Stream 2
Density	83	91
T65%	-	68
T85%	-	[a]
T95%	[a]	[a]
Cetane Number	90	[a]
Viscosity	-	33
Flash Point	[a]	[a]
Cloud Point	[a]	[a]
CFPP	-	[a]
Sulfur content	[a]	[a]
FAME content	-	96

[a] Models with R^2-values < 0.75 were not considered.

Partial conclusions

In this section, the stability of the PLS models developed for all properties was monitored over 10 months using samples that include new sources of variability. For most properties, except flash point (stream 2), sulfur content (stream 1), and FAME content, the results suggest that the models retained, to some extent, their stability over the period of time considered. Though, despite the observed deterioration of the predictive ability of the PLS model for FAME content over this 10-month period, it passed the test to verify agreement with the reference method and was the only model that remained valid for control in routine diesel analysis.

In the context of routine diesel analysis, despite the observed stability of the models, monthly updating of those models that exhibit good predictive ability is a requirement for their continued use. Novel diesel samples with Q and T^2 values beyond the limits of applicability must be added to the calibration dataset and the models should be recalculated to ensure robustness.

VI.3 Calibration transfer of a PLS model between instruments

A practical constraint in routine diesel analysis occurs when the model is applied to spectra acquired in a spectrophotometer that is different from the one where the model was developed. This often leads to poor predictive performance on new samples, rendering the model invalid under new conditions. This situation is considered in the present section.

To overcome the loss of model performance, the calibration model can be transferred to another instrument by using some calibration-model adaptations (sometimes referred to as model maintenance, model update and calibration transfer) [3–5]. The calibration-model adaptations have commonly been compared in terms of their respective RMSEP values. However, the improvement of RMSEP does not guarantee that the values estimated after the adaptation of the calibration model agree with the values of the reference method. Based on ASTM-E-1655 [6], the tolerance limits admitted by the method are defined from the reproducibility of the reference method ($\pm$R). In order to consider the model adaptation valid for routine analysis, not only the RMSEP value but also the percentage of samples whose prediction error falls within the tolerance limits admitted by the reference method must be compared. In some cases, the precision of the method is very high (that is, very low interlaboratory variability), and the tolerance limits to control the difference between the estimated values and the reference values can be expanded to $\pm$2R. In this way, although the model can not strictly replace the reference method, it can provide predictions that are valid enough to be used to control for the production process.

The aim of this section is to compare three approaches for the adaptation of an existing PLS calibration model based on IR spectroscopy for diesel analysis according to ASTM-E-1655. To this end, the suitability of unsupervised Domain invariant PLS (di-PLS), supervised Dynamic Orthogonal Projection (DOP), and Model Updating (MU) to correct for external influences associated with two spectrophotometers (instrumental drift) in predicting a quality parameter of diesel samples is compared for the first time. The comparison was carried out considering the RMSEP values and the percentage of samples with a prediction error within the tolerance limits of the reference method according to ASTM-E-1655.

From the practical point of view, the supervised DOP and MU require both spectra and reference values in the new conditions, while the unsupervised di-PLS only requires new spectra. These methods are especially relevant not only for petrochemical analysis

but also for industrial applications where the provision of reference values and complete recalibration are cumbersome. Furthermore, the results of the comparative study promote the widespread use of the PLS model for diesel analysis among different laboratories.

Materials and methods

The existing PLS calibration model for density in diesel was built with 1156 samples produced at the Repsol refinery in Tarragona (Spain) from 2018 to 2022. The reference density values of the samples were measured following the ASTM D4052 method [4] with an ANTON PAAR digital densimeter model DMA 4500M in the analysis laboratory. The spectra were measured using the procedure described in section III.4.1 of Chapter III.

A new Analect Diamond 20 FTIR/FT-NIR spectrophotometer with the same instrumental settings was purchased to back up the first one, and a new batch of 105 samples was measured on it. This will be called the target spectrophotometer. The determination of density in the new samples from their infrared spectra using the current calibration PLS model produced numerous outlier warnings. Therefore, a transfer of the existing model to the target spectrophotometer became necessary.

The non-linear iterative partial least squares (NIPALS) algorithm was used to develop the PLS model. The spectra were pretreated with Savitzky–Golay's first derivative (second-order polynomial and a 15-point window) and mean-centered. The optimal spectral region for density was a joint NIR (5168.29-3270.68 cm^{-1}) and MIR (2742.28-1801.19 cm^{-1}) region, so the PLS model was based on NIR and MIR spectroscopies. The optimal number of LVs of the model was selected based on the lowest prediction error of 10-fold cross-validation. According to the ASTM-E-1655, the model was considered to be validated for routine use standard if model predictions agree with the reference method using the procedure described in section II.3.2 of Chapter II.

The predictive ability of the model for the target condition was evaluated by RMSEP, and the determination coefficient of prediction (R_p^2) obtained after regressing the predicted versus the reference values. When the model was detected to be invalid due to changes in the instrument response, three approaches were tested for the adaptation of the existing PLS calibration model to the new situation. These approaches, domain invariant partial least squares unsupervised di-PLS, dynamic orthogonal projection DOP,

and model updating MU can lead to more general models that allow the transfer of the calibration models developed from the source to the target spectrophotometer.

Di-PLS consists of an extended PLS regression with a domain regularization term in order to align the spectra distributions on source and target spectrophotometers in the latent space [8]. To do this, we used a subset of the spectra of the new batch measured in the target spectrophotometer (which will be referred to as the transfer samples) to minimize the variability between the source and the target data matrix. The pretreatment used for di-PLS was column mean-centering, and the optimal number of LVs was obtained by external validation with a test set of 90 samples.

DOP involves recalculating the model by correcting the calibration spectra used for the existing PLS model. This correction is achieved through the calculation of virtual standards of transfer samples to obtain the orthogonalization matrix that filters the differences between the source calibration spectra and the target spectra [9]. In this case, the spectra are also mean-centered, and the optimal number of LVs was obtained by cross-validation.

MU involves recalculating the model by including the spectra of transfer samples measured on the target spectrophotometer into the current calibration set, which should reflect the spectral variability of the samples measured on two spectrophotometers [10]. The spectral preprocessing and the selection of LVs of the MU were the same as for the existing PLS model.

The external validation of all adaptation models was carried out by predicting the samples of the new batch that had not been used for the transfer. More details concerning the adaptation models can be found elsewhere [8,9,11–17].

The PLS models were developed using PLS-Toolbox v7 (The Eigen Vector Research, Manson, WA) in MATLAB (The MathWorks Inc., Natick, MA, R2022a), and the calculations for calibration transfer were carried out in MATLAB with homemade routines.

Data sets

The data set used to develop the existing model PLS in the laboratory of Repsol was randomly split into a training set (766 samples) and a validation set (390 samples), which contain 75% and 25% of the samples analyzed from both streams each month over 35 months, respectively. A new batch of 105 samples was used in both spectrophotometers

to develop the adaptation of the existing PLS calibration model to the target spectrophotometer. From this set, the samples for transferring and updating the model were selected with the Kennard-Stone algorithm [18]. As a result, the new batch was divided into a transferring set (15 samples) used to transfer the calibration model and a test set (90 samples) used to perform the external validation.

The density values of all samples were determined using the reference method mentioned above. The density values ranged from 820.0 kg/m^3 to 845.0 kg/m^3. A summary of the data sets involved in the calibration and adaptation models is provided in Table VI-3 and Table VI-4, respectively.

Table VI-3. Summary of the datasets used in the PLS model.

Data set	PLS	
	Calibration (samples × wavelengths)	Validation (samples × wavelengths)
Spectra	$\mathbf{X}_{\mathrm{cal_S}}$ (766×738) (S)	$\mathbf{X}_{\mathrm{val_S}}$ (390×738) (S)
Reference values	$\mathbf{y}_{\mathrm{cal_S}}(766 \times 1)$	$\mathbf{y}_{\mathrm{val_S}}(390 \times 1)$

(S) source spectrophotometer and (T) target spectrophotometer

Table VI-4. Summary of the datasets involved in the adaptation of the PLS model.

Data set	Transferred PLS (di-PLS and DOP)		Model updating	
	Transfer (samples × wavelengths)	Test (samples × wavelengths)	Updating (samples × wavelengths)	Test (samples × wavelengths)
Spectra	$\mathbf{X}_{\mathrm{trans_T}}(15 \times 738)$ (T)	$\mathbf{X}_{\mathrm{test_T}}(90 \times 738)$ (T)	$\mathbf{X}_{\mathrm{cal_S}}(766 \times 738)$ + $\mathbf{X}_{\mathrm{trans_T}}(15 \times 738)$ (S) +(T)	$\mathbf{X}_{\mathrm{test_T}}(90 \times 738)$ (S) +(T)
Reference values	$\mathbf{y}_{\mathrm{trans_T}}(15 \times 1)$	$\mathbf{y}_{\mathrm{trans_T}}(90 \times 1)$	$\mathbf{y}_{\mathrm{cal_S}}(766 \times 1)$ + $\mathbf{y}_{\mathrm{trans_T}}(15 \times 1)$	$\mathbf{y}_{\mathrm{trans_T}}(90 \times 1)$

(S) source spectrophotometer and (T) target spectrophotometer

Results and discussion

Evidencing the need for model adaptation

The existing PLS model was constructed using 12 LVs determined by cross-validation that accounted for 98.97% and 99.17% of the $\mathbf{X}_{\mathrm{cals}}$ and $\mathbf{y}_{\mathrm{cals}}$ variance, respectively. The RMSEP and the $\mathrm{R}_{\mathrm{p}}^2$ of the linear fit between the reference density and the predicted values for the validation set in the source spectrophotometer were $0.50\ \mathrm{kg/m^3}$ and 0.99, respectively. These values are comparable to previously published models developed using spectroscopic techniques [8,22,23]. In addition, based on the referred ASTM-E-1655 [6], the values predicted by the PLS calibration models agreed with the reference method since 95% of the prediction errors of the validation set fell in the range $\pm 2\mathrm{R}$ (Figure VI-4). Thus, the model was considered valid for routine analysis.

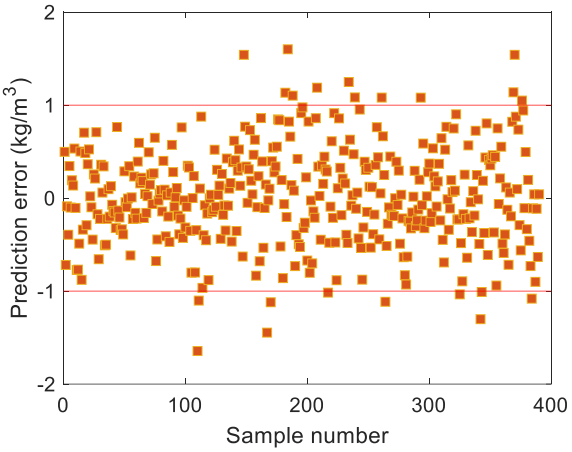


Figure VI-4. Prediction error vs. sample number of the new batch.

Before applying the PLS model, it is necessary to ensure that the new samples are similar to the calibration and validation samples used for the establishment of the model. The limit of applicability of the PLS calibration model can be used to address this fact. Hotelling's T^2 and Q-residuals statistics as a measure of leverage and spectral residuals, respectively, are commonly used to identify new spectral variability in new spectra. The applicability limits $\mathrm{T}_{\mathrm{lim}}^2$ and $\mathrm{Q}_{\mathrm{lim}}$, can be set as a certain quantile (e.g., 0.95 or 0.99) of the distribution that follows the Q or T^2 values of the training dataset. New samples that produce Q or T^2 values larger than $\mathrm{Q}_{\mathrm{lim}}$ and $\mathrm{T}_{\mathrm{lim}}^2$ do not necessarily imply that the predictions are flawed but rather that they should not be trusted [3]. Hence, the predictions should be verified against the reference method.

When the spectra of the new batch were measured in the target spectrophotometer, the first task was to check the spectral disturbances due to changes in the instrument response. The raw absorbance spectra acquired from the source and target

spectrophotometer are presented in Figure VI-S1 in the Supplementary Information. The first task was to check whether the new spectra were well represented by those used to establish the model. To do this, Figure VI-5 shows the first derivative of the spectra in the two spectral regions that were used for calibration. In both regions, differences remain in the spectral pattern that are not even removed by the first derivative. Although the spectra shown in Figure VI-5 are the result of measuring different samples in both spectrophotometers, it can be seen that the main differences are horizontal shifts of the absorption bands in both regions. In particular, a small wavenumber shift occurs around the absorption bands of 4578, 4184, 3814, 3764, 3721, and 3594 cm^{-1}.

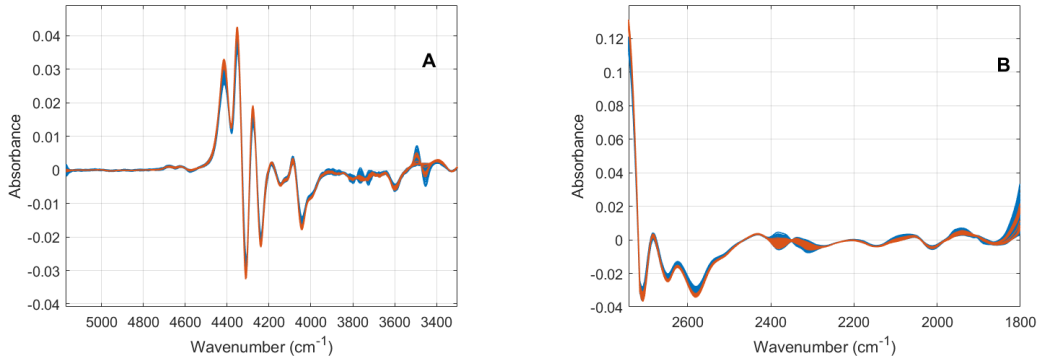


Figure VI-5. First derivative spectra of the diesel samples acquired on the source (blue) and target spectrophotometer (orange) in A) the NIR region and B) the MIR region.

As expected, the Q-residuals versus Hotelling's T^2 (Figure VI-6A) revealed the abnormal behavior of the new batch located in the upper right quadrant concerning the calibration and validation samples, since its Q and T^2 values exceeded the limits of applicability of the PLS model. The new spectra included new variability not considered during calibration and thus, a loss of predictive ability of the model under the new conditions can be expected. To verify this, the reference values corresponding to the new spectra were determined using the reference method. When the model was applied to predict the density values of these spectra, the predictive performance was poor with a RMSEP of $5.27\ kg/m^3$ (Figure VI-6B). Also, Figure VI-7 shows that the prediction errors for all the samples were higher than the expanded tolerance limits ($\pm 1\ kg/m^3$). Thus, calibration transfer methods and model updating were tested to overcome these spectral perturbations, and the loss of predictive ability of the model under the target instrument, calibration transfer methods and model updating were tested.

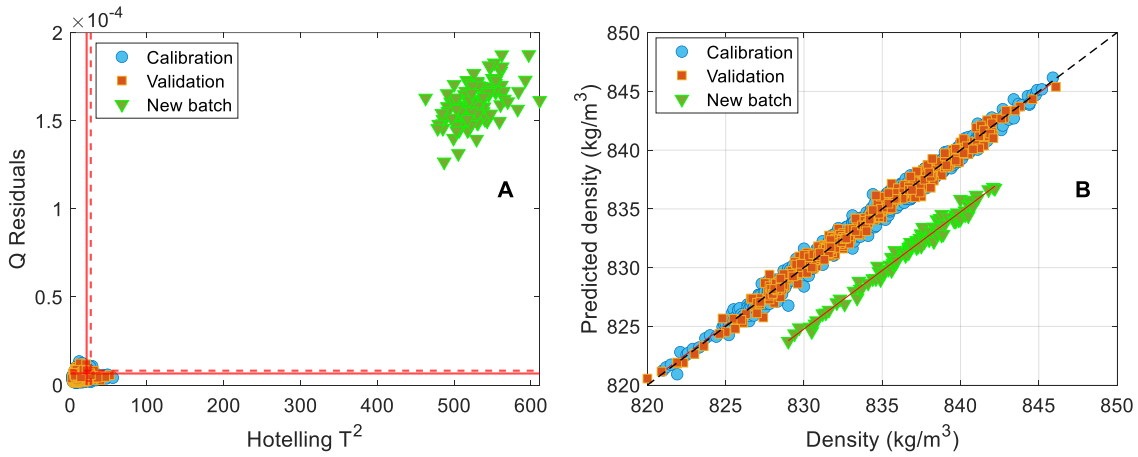


Figure VI-6. A) Q-residuals vs. Hotelling's T^2. B) Predicted vs measured values of density with the current model. Calibration samples of the established models (blue), validation samples of the established models (orange), and samples of the new batch (green).

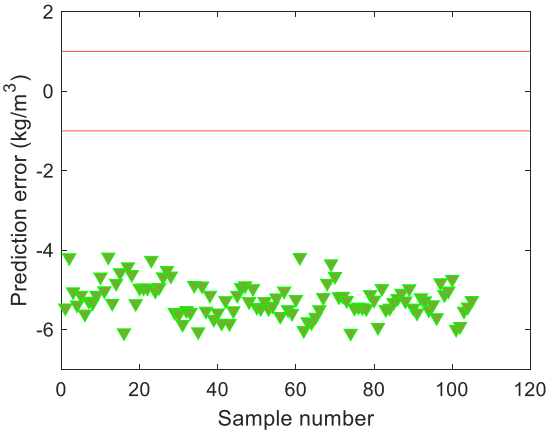


Figure VI-7. Prediction error vs. sample number of the new batch.

Instrument adaptation models to predict the density of diesel samples

Di-PLS regression (Figure VI-8A) with 7 LVs using only 15 transfer samples representative of the variations in target data improved the RMSEP obtained for validation samples to $1.04\ \mathrm{kg/m^3}$, which is close to the tolerance limits admitted by the reference method. The prediction error for the target instrument decreased by 80%. The improvement in terms of the predictive ability of the di-PLS model was attributed to the alignment of the distributions of the scores corresponding to $\mathbf{X}_{\mathrm{cal\_S}}$ and $\mathbf{X}_{\mathrm{trans\_T}}$, which indicates that the new space of latent variables is invariant concerning the distributional properties of the domains of the spectra measured on the source and target spectrophotometers. This alignment led to a lower prediction error than the prediction error of the PLS developed with data from the source instrument only. However, the $\mathrm{R_p^2}$ was 0.87, which means the $\mathrm{R_p^2}$ decreased by 12%, indicating that the di-PLS did not

successfully recover the functional relationship between IR spectra and the density of samples. In addition, it can be observed in Figure VI-8B that a significant number of samples are outside the interval determined by the admitted tolerance limits. Overall, di-PLS calibration transfer was not a suitable strategy to overcome the loss of the predictive ability of the PLS model with the target spectrophotometer.

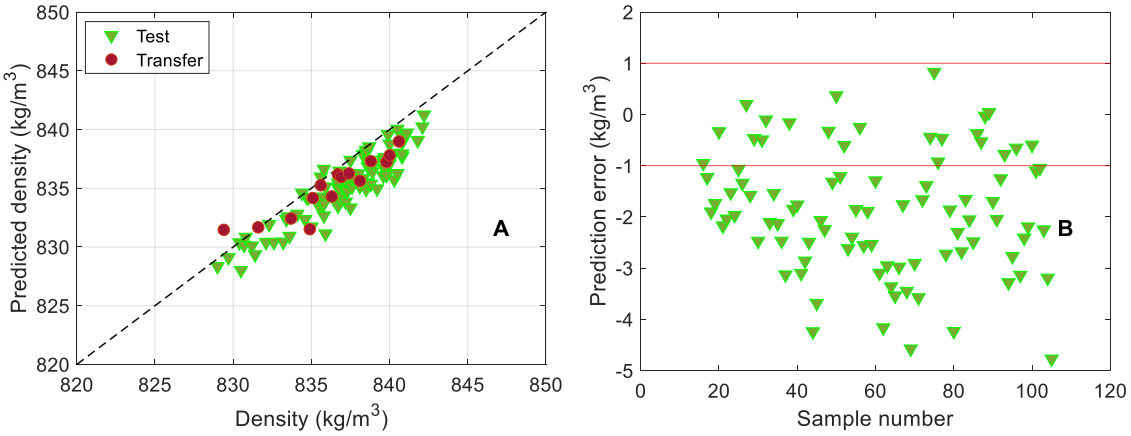


Figure VI-8. A) Predicted density vs. measured density after di-PLS. B) Prediction error after di-PLS vs. sample number.

After DOP correction (Figure VI-9A), 16 LVs were required to achieve a model with the lowest cross-validation error. The value of R_p^2 for validation samples was 0.99, while the average prediction error was $\mathrm{RMSEP} = 0.69\ \mathrm{kg/m^3}$. The DOP correction improved the performance of the model since the prediction error decreased by 87%, and the coefficient of determination remained unchanged. These results indicated the superior predictive ability with respect to the original PLS and also di-PLS. In this case, the correction applied to the source calibration spectra through orthogonal projection is integrated into the model itself. As a result, there is no need to correct new spectra acquired with the target spectrophotometer when using the recalculated model. Although the number of validation samples that are outside the admitted tolerance limits (Figure VI-9B) has decreased significantly compared to the existing PLS model (10% of samples) the model was not yet valid for routine analysis according to ASTM-E-1655.

MU involved 15 LVs, three LVs more than in the original model to include the new variability introduced by the measurement in the target spectrophotometer. As expected (Figure VI-10A), the RMSEP decreased by 92% ($\mathrm{RMSEP} = 0.44\ \mathrm{kg/m^3}$), which indicates the superior predictive ability with respect to the original PLS and also transferred models (di-PLS and DOP). In this case, only 3% of the prediction error of the validation samples

falls out of the tolerance limits admitted (Figure VI-10B). Thus, the MU agrees with the reference method and is valid for control in routine diesel analysis. Table VI-5 summarizes all the above results.

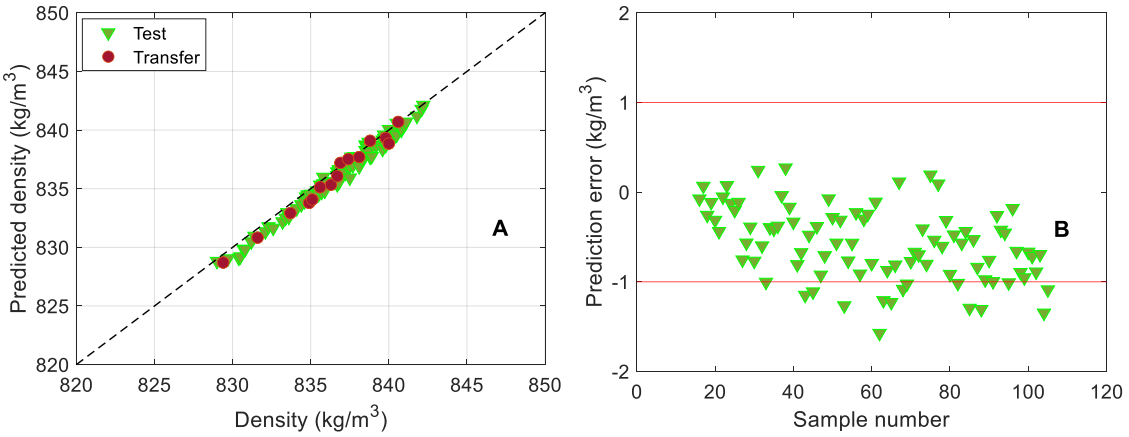


Figure VI-9. A) Predicted density vs. measured density after DOP correction. B) Prediction error after DOP vs. sample number.

It is important to mention that no more than 15 transfer samples were needed to improve the results since the number of samples used was the minimum necessary to minimize the RMSEP of the external validation set. By increasing the number of transfer samples (from 15 to 70), the RMSEP for the external validation set did not improve significantly.

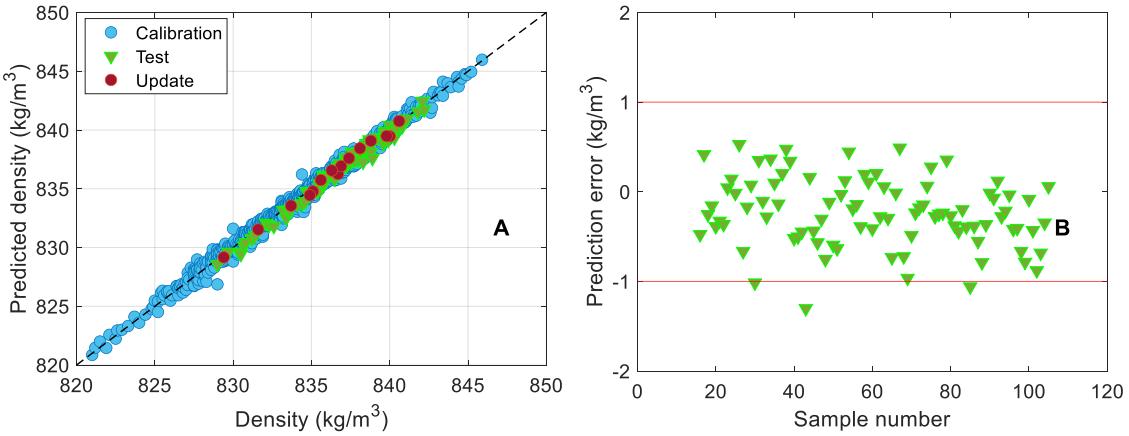


Figure VI-10. A) Predicted density vs. measured density after MU. B) Prediction error after MU vs. sample number.

Table VI-5. Prediction of density by PLS, di-PLS, DOP, and MU.

Parameters	Existing model	Transferred models		Updated model
	PLS	di-PLS	DOP	MU
$\mathbf{N_v}$	390	90		
LVs	12	7	16	15
$\mathbf{R_p^2}$	0.99	0.87	0.99	0.99
RMSEP	0.52	1.04	0.69	0.44

Partial conclusions

The PLS regression model used to determine density in diesel samples based on NIR and MIR spectroscopies showed a good predictive ability. However, when the model was applied to spectra acquired in a new spectrophotometer, the predictive ability was poor, and the model was not valid for routine analysis.

The three considered adaptations of the calibration model improved the predictive ability of the model in the new conditions for the determination of density with reductions of 80%, 87%, and 92% in RMSEP after applying di-PLS, DOP and MU, respectively. However, only MU passed the test to verify agreement with the reference method and, hence, was considered valid for control in routine diesel analysis.

VI.4 Conclusions

Although monitoring the predictive ability of the PLS models for ten months showed that the prediction errors did not vary significantly for most of the properties, only the PLS model for predicting FAME content remained valid for routine diesel analysis.

Three model adaptation techniques were explored to address deteriorating predictive ability when the PLS model for density was applied to spectra acquired in a new spectrophotometer. MU proved to be the most effective of the three approaches (92% reductions in RMSEP). MU is the only tested model adaptation that can be considered valid for routine diesel analysis since only the 3% of the validation samples falls out of the tolerance limits admitted by the reference method.

The present comparative study provides valuable information to the audience interested in infrared technology as it addresses possible solutions to changes in instrumental response when using different spectrophotometers in calibration and prediction steps.

The findings presented herein may be of interest in industrial applications where analytical methods based on multivariate calibration and infrared data are used in routine analysis, such as in the food and pharmaceutical industries.

VI.5 References

[1] X. Capron, B. Walczak, O.E. De Noord, D.L. Massart, Selection and weighting of samples in multivariate regression model updating, Chemom. Intell. Lab. Syst. 76 (2005) 205–214. https://doi.org/10.1016/j.chemolab.2004.11.003.

[2] J.B. Cooper, C.M. Larkin, M.F. Abdelkader, Calibration transfer of near-IR partial least squares property models of fuels using virtual standards, J. Chemom. 25 (2011) 496–505. https://doi.org/10.1002/cem.1395.

[3] B.M. Wise, R.T. Roginski, A calibration model maintenance roadmap, IFAC-PapersOnLine. 28 (2015) 260–265. https://doi.org/10.1016/j.ifacol.2015.08.191.

[4] R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown, J. Ferré, Transfer of Multivariate Calibration Models, Chemom. Intell. Lab. Syst. 64 (2002) 181–192. https://doi.org/10.1016/B978-044452701-1.00077-6.

[5] J.J. Workman, A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy, Appl. Spectrosc. 72 (2018) 340–365.

https://doi.org/10.1177/0003702817736064.

[6] American Society for Testing Materials, ASTM E1655-17 Standard Practices for Infrared Multivariate Quantitative Analysis, 2017. https://doi.org/10.1520/E1655-17.

[7] D40052-15, Standard Test Method for Density , Relative Density , and API Gravity of Liquids by Digital Density Meter, ASTM Int. (2013) 1–8. https://doi.org/10.1520/D4052-18A.2.

[8] R. Nikzad-Langerodi, W. Zellinger, E. Lughofer, S. Saminger-Platz, Domain-Invariant Partial-Least-Squares Regression, Anal. Chem. 90 (2018) 6693–6701. https://doi.org/10.1021/acs.analchem.8b00498.

[9] M. Zeaiter, J.M. Roger, V. Bellon-Maurel, Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations, Chemom. Intell. Lab. Syst. 80 (2006) 227–235. https://doi.org/10.1016/j.chemolab.2005.06.011.

[10] M.R. Kunz, J.H. Kalivas, E. Andries, Model updating for spectral calibration maintenance and transfer using 1-norm variants of tikhonov regularization, Anal. Chem. 82 (2010) 3642–3649. https://doi.org/10.1021/ac902881m.

[11] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, B. Moser, Domain-invariant regression under beer-lambert's law, Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019. (2019) 581–586. https://doi.org/10.1109/ICMLA.2019.00108.

[12] R. Nikzad-Langerodi, W. Zellinger, S. Saminger-Platz, B.A. Moser, Domain adaptation for regression under Beer–Lambert's law, Knowledge-Based Syst. 210 (2020) 106447. https://doi.org/10.1016/j.knosys.2020.106447.

[13] P. Mishra, R. Nikzad-Langerodi, Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit, Infrared Phys. Technol. 111 (2020) 103547. https://doi.org/10.1016/j.infrared.2020.103547.

[14] B. Mikulaseka, V. Fonseca Diaz, C. Herwig, D. Gabauer, R. Nikzad-Langerodi, Partial Least Squares Regression With Multiple Domains, J. Chemom. 37 (2023) 3477. https://doi.org/10.1002/cem.3477.

[15] P. Mishra, J.M. Roger, D.N. Rutledge, E. Woltering, Two standard-free approaches to correct for external influences on near-infrared spectra to make models widely applicable, Postharvest Biol. Technol. 170 (2020) 111326.

https://doi.org/10.1016/j.postharvbio.2020.111326.

[16] P. Mishra, J.M. Roger, F. Marini, A. Biancolillo, D.N. Rutledge, FRUITNIR-GUI: A graphical user interface for correcting external influences in multi-batch near infrared experiments related to fruit quality prediction, Postharvest Biol. Technol. 175 (2021) 111414. https://doi.org/10.1016/j.postharvbio.2020.111414.

[17] P. Mishra, CT-GUI: A graphical user interface to perform calibration transfer for multivariate calibrations, Chemom. Intell. Lab. Syst. 214 (2021) 104338. https://doi.org/10.1016/j.chemolab.2021.104338.

[18] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.

[19] V.O. Santos Jr., 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. 547 (2005) 188–196. https://doi.org/10.1016/j.aca.2005.05.042.

[20] M.G. Nespeca, R.R. Hatanaka, D.L. Flumignan, E. De Oliveira, Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis, J. Anal. Methods Chem. (2018) 1795624. https://doi.org/10.1155/2018/1795624.

[21] S. Marinović, M. Krištović, B. Špehar, V. Rukavina, A. Jukić, Prediction of diesel fuel properties by vibrational spectroscopy using multivariate analysis, J. Anal. Chem. 67 (2012) 939–949. https://doi.org/10.1134/S1061934812120039.

VI.6 Supplementary Information

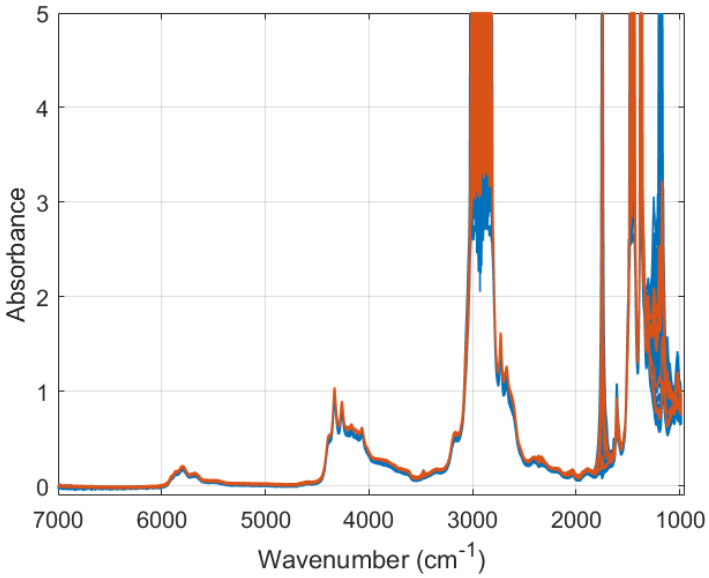


Figure VI-S1. Raw spectra of diesel samples acquired from the source (blue) and target spectrophotometers (red).

Chapter VII

Conclusions and perspectives

VII.1 Conclusions

The development of multivariate calibration models for the determination of quality properties of diesel from its IR spectrum was addressed in this thesis. PLS- and ANN-based regressions have been able to model the complex IR spectrum–property relationships in diesel samples. Two diesel streams, namely desulfurized diesel (stream 1) and commercial diesel (stream 2), and eleven physiochemical properties of diesel, namely temperatures at 65%, 85%, and 95% recovered, flash point, cloud point, density, cetane number, sulfur content, viscosity, cold filter plugging point (CFPP), fatty acid methyl esters (FAME) have been studied.

The main findings of this thesis lead to the following conclusions:

1. The analysis of diesel quality parameters in the samples analyzed evidenced that:
   - Desulfurized diesel samples were characterized by higher T95%, flash point, cloud point, density, cetane number, and sulfur content values than commercial diesel samples.
   - Regardless of the sample origin (desulfurized or commercial samples), correlations between property values were generally below 0.6, except for the distillation temperatures of commercial samples, which exhibited correlations as high as 0.9. These results suggested that the correlations between properties showed a weak dependence on the chemical composition of the samples.

2. Exploratory analysis of IR spectra based on classical and multivariate spectral analysis showed that:
   - In the spectral region considered 7000-950 cm^{-1}, the specific band of the carbonyl group at 1750 cm^{-1} enabled the rapid identification of samples containing FAME.
   - PCA of the spectra differentiated samples into two clusters based on their FAME content. t-SNE successfully differentiated desulfurized samples, commercial samples containing FAME, and commercial samples without FAME content. The difference in the property values of the samples seems to be more influenced by the origin of the samples (streams 1 or 2) rather than by the FAME content.

3. It was possible to predict accurately the density of desulfurized and commercial diesel fuel samples from their MIR spectra using a global calibration model based on feed-forward neural network (FFNN) regression.

4. A new methodology was developed to define the applicability limits of an FFNN model that allowed the discarding of erroneous spectra measured in a second instrument and discordant spectra. The limits were defined from 0.99 quantiles of two metrics: 1) the squared Mahalanobis distance calculated from the activations of the hidden layer of the FFNN, and 2) the sum of squared spectral residuals obtained by reconstructing the spectrum from the original spectra (autoencoder approach) and from a lower dimensional space (decoder approach).

5. The hypothesis that the decoder might outperform the autoencoder could not be confirmed. The ability to reproduce the prediction spectrum from the training data for the autoencoder and decoder approaches was similar.

6. Multivariate calibration model based on partial least squares (PLS) and non-linear approaches based on FFNN yielded accurate predictions for properties closely related to the chemical composition of diesel: density, cetane number, viscosity, and content from FAME. Therefore, these analytical methods based on IR spectroscopy coupled with either of the two calibration models can be considered as an alternative to standard methods. The selection of wavelengths is relevant to optimize the predictive ability of each model.

7. The sulfur content predictions obtained by both multivariate models were poor, probably due to the low concentration present in the analyzed samples and the low molar absorptivity of the sulfur compounds in the infrared region.

8. The PLS and ANN calibration models yielded poor predictions for properties not directly related to the chemical composition of diesel: distillation temperatures such as T65%, T85%, and T95%, flash point, cloud point, and CFPP. Thus, neither model can be considered an alternative to the ASTM reference method for diesel quality analysis. However, both calibration models did provide reasonable estimates for most of these properties (T65%, T85%, T95%, flash point, and cloud point). For these properties, the developed models might prove valuable in control processes with less stringent accuracy demands for fast detection of gross deviations of normal process conditions.

9. ANN models were superior to PLS models for predicting density, T65%, T85%, flash point, CFPP, and FAME content in commercial samples, and T95%, flash point, cloud point, sulfur content in desulfurized samples. In contrast, PLS predicted better density in desulfurized samples, T95%, cloud point, and sulfur content in commercial samples, and cetane number in desulfurized and commercial samples. Moreover, the

ANN and PLS models showed similar predictive ability for viscosity in commercial samples. Overall, the predictive ability of both PLS and ANN models for properties in commercial samples was superior, with the exception of the cetane number in desulfurized samples.

10. For most properties, -except flash point (commercial samples), sulfur content (desulfurized samples), and FAME content- the developed PLS models were stable over a 10-month monitoring period. Nevertheless, despite the observed deterioration of the predictive ability of the PLS model for FAME content over this monitoring period, it was the only one that remained valid for the control in routine diesel analysis.

11. The PLS calibration model used to determine the density in commercial diesel samples was no longer valid for predicting this property when applied to spectra acquired in a new spectrophotometer with similar characteristics to the one used during model development.

12. Three calibration model adaptations, di-PLS, DOP, and MU, were considered to transfer an existing PLS model to the new instrument, improving its predictive ability. Only the adaptation based on MU yielded valid results for routine diesel analysis control.

VII.2 Perspectives

A prospective path for further research involves exploring how combining infrared measurements with complementary analytical techniques or with chemical composition data could enhance the performance of calibration models for predicting the physicochemical properties of diesel. For instance, considering the concentration of additives like cetane improvers and cold-flow improvers in diesel samples could improve the predictive ability of calibration models for cetane number and low-temperature properties, respectively.

While this study has focused on FFNN regression models, it is worth considering other different types of neural networks, such as Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs), Radial Basis Function Networks (RBFNs), and so on, that offer different architectural features and learning mechanisms.

As an alternative to ANN, support vector machines (SVMs) could be a promising option for spectral regression tasks, especially in models where poor predictions were obtained for properties not directly linked to the chemical composition of diesel.

The approach used to adapt an existing PLS model based on the optimal NIR (5168.29-3270.68 cm^{-1}) and MIR (2742.28-1801.19 cm^{-1}) spectral region for density determination to a new instrument could similarly be employed to adjust two PLS models based on each of these spectral subregions. The development and comparison of these two models could provide valuable insights into their robustness and the sensitivity of each spectral region to instrumental variations.

Improvements in the predictive ability of the models for cetane number and viscosity that were based on a small number of samples could be expected as new diesel spectra are measured and their reference property values are determined and incorporated into the models.

Regarding the calibration transfer approaches, this study has focused on the evaluation of three models adapted from PLS for predicting density in diesel samples, but models based on transfer learning (TL) for adapting artificial neural networks (ANN) models could also be considered in future studies.

UNIVERSITAT
ROVIRA i VIRGILI