# Digital process monitoring par l'analyse en ligne de l'effluent total en proche infrarouge pour déterminer les qualités produits de coupes

Jhon Buendia Garcia

# THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'INSTITUT AGRO MONTPELLIER ET DE L'UNIVERSITE DE MONTPELLIER

**En Génie des procédés**

**École doctorale : GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau**

**Portée par**

**IFP Energies nouvelles à Solaize, et de l'unité mixte de recherche Information, Technologie, Agro-Procédés (ITAP) à Montpellier**

# DIGITAL PROCESS MONITORING PAR L'ANALYSE EN LIGNE DE L'EFFLUENT TOTAL EN PROCHE INFRAROUGE POUR DÉTERMINER LES QUALITÉS PRODUITS DE COUPES

**Présentée par Jhon BUENDIA GARCIA**

**Le 12 Juillet 2022**

**Sous la direction de Dr. Jean-Michel ROGER et Dr. Ryad BENDOULA**

**Devant le jury composé de**

| | |
|---|---|
| Marina COCCHI, Professeure associée, Université UNIMORE, Italie | [Rapportrice] |
| Nida, SHEIBAT-OTHMAN, Professeure, Université Claude Bernard Lyon 1, France | [Rapportrice] |
| Douglas N. RUTLEDGE, Professeur émérite, AgroParis Tech, France | [Examinateur-président du jury] |
| Nadège BRUN, Ingénieure de recherche, TotalEnergies, France | [Examinatrice] |
| Julien GORNAY, Ingénieure de recherche, IFPEN, France | [Examinateur] |
| Ryad BENDOULA, Directeur de recherche, INRAE, Université de Montpellier, France | [Directeur] |
| Jean-Michel ROGER, ICPEF, INRAE, Université de Montpellier, France | [Invité] |
| Marion LACOUE-NEGRE, Ingénieure de recherche, IFPEN, France | [Invité] |

UNIVERSITÉ DE MONTPELLIER

L'INSTITUT agro Montpellier

# THESIS TO OBTAIN THE DEGREE OF DOCTOR OF PHILOSOPHY OF THE MONTPELLIER AGRO INSTUTE AND THE UNIVERSITY OF MONTPELLIER

**In Process Engineering**

**Doctoral School : GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau]**

**Supported by**
**IFP Energies nouvelles at Solaize, and the mixed research unit of Information, Technology, Agro-Processes (ITAP) at Montpellier**

# DIGITAL PROCESS MONITORING BY ONLINE NEAR-INFRARED ANALYSIS OF THE TOTAL EFFLUENT TO DETERMINE THE PRODUCT QUALITIES OF CUTS

**Submitted by Jhon BUENDIA GARCIA**
**Le 12 July 2022**

**Under the direction of Dr. Jean-Michel ROGER and Dr. Ryad BENDOULA**

**In front of the jury composed by**

| | |
|---|---|
| Marina COCCHI, Professeure associée, Université UNIMORE, Italie | [Reviewer] |
| Nida, SHEIBAT-OTHMAN, Professeure, Université Claude Bernard Lyon 1, France | [Reviewer] |
| Douglas N. RUTLEDGE, Professeur émérite, AgroParis Tech, France | [Referee-president of the jury] |
| Nadège BRUN, Ingénieure de recherche, TotalEnergies, France | [Referee] |
| Julien GORNAY, Ingénieure de recherche, IFPEN, France | [Referee] |
| Ryad BENDOULA, Directeur de recherche, INRAE, Université de Montpellier, France | [Director] |
| Jean-Michel ROGER, ICPEF, INRAE, Université de Montpellier, France | [Invited] |
| Marion LACOUE-NEGRE, Ingénieure de recherche, IFPEN, France | [Invited] |

UNIVERSITÉ DE MONTPELLIER

L'INSTITUT agro Montpellier

# Preface

The research conducted in this thesis was motivated by the need to improve the experimental efficiency in the research and development of the hydrocracking process (HCK). For this purpose, the work sought a methodology to estimate the properties of the middle distillates (diesel and kerosene) without performing the distillation of the total effluent obtained from the process reactors or the analysis of the physical cuts.

The development and implementation of the thesis were possible due to the collaboration between the research and training institute in the fields of energy, transport and environment IFP Energies Nouvelles (IFPEN) in Solaize (France), and the mixed research unit Informations, Technologies for Agro-Processes (ITAP) in Montpellier (France). The work was financially supported by IFPEN under the direction of Dr. Jean-Michel ROGER and Dr. Ryad BENDOULA, and under the supervision of Dr. Marion LACOUE-NEGRE, Dr. Julien GORNAY, and Dr. Silvia MAS GARCIA.

During the three years of research dedicated to the development of the thesis, the synergic and collaborative work of all participants involved resulted in novel and promising results not only for optimizing the experimental activity in HCK process research but also for the analysis and monitoring of this process.

Across the six chapters of the document, the reader shall find the sequence of the thesis development coherently and logically. The first chapter presents the context of the thesis, the motivation for the HCK process research, and the thesis research questions. Chapter 2 presents the materials and methods used to develop the thesis. Chapters 3 - 5 show the thesis development and the results that helped address the research questions formulated, while Chapter 6 reports the application of the findings in two cases study. Finally, the document presents the conclusions and perspectives of the research work.

The thesis presented in this document is submitted to obtain the Doctor of Philosophy (Ph.D.) title from the doctoral school GAIA of the University of Montpellier and the Montpellier Agro institute.

A aquellas personas que han estado siempre a mi lado: mis padres, mis hermanos,
y mi "black diamond" cara mia. Solo a Dios sea la gloria.

# Acknowledgments

First of all, I would like to thank the reviewers, Prof. Marina Cocchi and Prof. Nida SHEIBAT-OTHMAN, and other members of the jury, Prof. Douglas N. RUTLEDGE and Dr. Nadège BRUN, for having gladly accepted to evaluate this work.

I want to thank my thesis directors, Dr. Jean-Michel ROGER and Dr. Ryad BENDOULA, and my supervisors, Dr. Julien GORNAY, Dr. Marion LACOUE-NEGRE, and Dr. Silvia Mas Garcia, for guiding me in the development of the thesis. Their input, support, advice, and teachings contributed to a thesis development full of learning amid an ideal scientific and personal environment. It was a formidable work team with a great scientific spirit, each one contributing with his experience and his quality as a person.

Thanks to the director of the experimentation intensification department at IFPEN, Denis Guillaume, the the department's scientific associate, Hervé Cauffriez, and the head of the department, Fabrice Giroudière, for their welcome. I also want to thank those who made the experimental work possible in an optimal environment. To the online analysis research team, Sebastian, Aurelie and Maud. Special thanks to Noémie CAILLOL and Axel One for their constant willingness to provide the instruments required for the experimental analyses. Thanks to Cedric and Alexia in the pilot unit operation for always having an unquestionable technical rigor and collaboration.

To the ChemHouse chemometrics group, I would like to express a special thanks. A research group open to share its knowledge for the sake of scientific development. Each person involved contributed a grain of sand to the thesis development, from alternative solutions on how to solve the research questions to ideas on how to present the results.

To each person who offered their sincere companionship and support in crucial moments, especially Maria y Raquel.

Finalmente, agradezco el apoyo que siempre tuve de mi familia, allí dando ánimo a lo largo de todo el trabajo. A la posa amara, Cari, fue la pieza fundamental de este logro, sin ella este sueño no hubiera sido posible.

Eager to continue exploring the world of chemometrics.

# Publications and communications

## 1. Papers in international peer-reviewed journals

1. J. Buendia Garcia, J. Gornay, M. Lacoue-Negre, S. Mas Garcia, J. Er-Rmyly, R. Bendoula, J.-M Roger, A novel analysis methodology for preprocessing methods effectiveness determination in reducing undesired spectral variability in near-infrared spectra acquisition. Journal of NIR Infrared Spectroscopy (JNIRS), Vol 30, issue 2, 2022. https://doi.org/10.1177/09670335211047959
2. J. Buendia Garcia, M. Lacoue-Negre, J. Gornay, S. Mas Garcia, R. Bendoula, J.-M Roger, Diesel cetane number estimation from NIR spectra of hydrocracking. (Submitted to FUEL and accepted)
3. J. Buendia Garcia, S. Mas Garcia, M. Lacoue-Negre, J. Gornay, R. Bendoula, J.-M Roger, NIR and 13C NMR data fusion to improve diesel cold flow properties prediction. (Submitted to Fuel)
4. 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. (to be submitted shortly)
5. J. Buendia Garcia, J.-M Roger, S. Mas Garcia, M. Lacoue-Negre, J. Gornay, R. Bendoula, Application of orthogonalization methods for robust diesel cetane number estimation from hydrocracking total effluent NIR spectra. (to be submitted shortly)

## 2. Oral communications

1. J. Buendia Garcia, M. Lacoue-Negre, J. Gornay, S. Mas Garcia, R. Bendoula, J.-M Roger, Diesel cetane number prediction by data fusion of near-infrared and nuclear magnetic resonance spectroscopy, in: e-chimiométrie 2021, France, 2021
2. J. Buendia Garcia, M. Lacoue-Negre, J. Gornay, S. Mas Garcia, R. Bendoula, J.-M Roger, Evaluation of variable selection methods and heterogeneous block data fusion strategies for improvement of diesel cetane number prediction, in: Road to CAC 2022, France - International, 2021

# Résumé étendu

Le basculement de la consommation de l'essence vers le diesel, l'augmentation de la production de pétrole brut lourd et la demande constante de produits de haute qualité ont fait émerger le besoin de procédés de raffinage flexibles qui maximisent la production de distillats moyens (kérosène et gazole) à partir de charges de plus en plus lourdes tout en garantissant leur qualité afin de satisfaire aux législations environnementales et commerciales. En raison de sa grande flexibilité dans le traitement des charges lourdes, le procédé d'hydrocraquage (HCK) est fondamental pour répondre au besoin décrit. Ce procédé est aujourd'hui largement mis en œuvre dans les raffineries. C'est pour cette raison qu'il fait l'objet de recherches permanentes.

La recherche sur le procédé HCK est notamment conduite en implémentant des plans expérimentaux dans des unités pilotes et des installations de laboratoire avec des conditions opératoires contrôlées. L'expérimentation ainsi mise en œuvre contribue à déterminer la meilleure configuration du procédé en traitant différentes charges types de résidus, principalement des distillats sous vide (DSV), à pour différentes conditions opératoires. Cette optimisation peut être décomposée en deux parties : En général, le déroulement des expériences se compose de deux étapes principales : (i) la section de réaction catalytique où se produisent les réactions d'hydrotraitement et d'hydrocraquage pour obtenir un effluent liquide léger appelé effluent total, et (ii) l'étape de distillation de l'effluent totalcet effluent et puis de caractérisation des coupes de distillationproduits. Les différentes coupes pétrolières générées produits HCK, en particulier les distillats moyens, sont  sont caractérisées en utilisant différentes méthodes normesnormalisées standard telles que celles de l'American Society for Testing and Materials (ASTM) et ou de l'Organisation internationale de normalisation (ISO) afin de vérifier notamment si elles respectent les spécifications des marchés. Contrairement à l'étape réactionnelle, la caractérisation des produits est effectuée de manière discontinue. Premièrement, les analyses de laboratoire sont effectuées hors ligne et sont conditionnées par les temps de réponse des différents laboratoires. Par ailleurs, pour effectuer les analyses de laboratoire en utilisant les normes mentionnées précédemment, l'échantillon physique du produit doit être obtenu à partir de la distillation de l'effluent total, qui est également effectuée dans une séquence non continue. La caractérisation des produits est une tâche fondamentale dans la recherche sur le procédé HCK. Cependant, le schéma analytique traditionnellement suivi exige à la fois du temps et du volume d'échantillon ce qui est contraignant et limite les temps de développement. Par conséquent, une alternative rapide, robuste et fiable pour la caractérisation des distillats moyens a été développée dans cette thèse.

La première étape de la thèse a été de proposer une alternative de caractérisation permettant de lever la contrainte de temps liée à la distillation de l'effluent total et à la caractérisation des produits obtenus. Dans la littérature, différents articles mentionnent que sur la base d'informations spectroscopiques acquises pour

les distillats moyens, les propriétés de ces produits ont été estimées avec des performances statistiques proches des méthodes de référence standard normalement utilisées. Cette alternative a permis de réduire à la fois le volume d'échantillon nécessaire et le temps de réponse de l'analyse. Cependant, s'agissant d'une estimation basée sur les informations analytiques acquises sur les coupes des distillats moyens, l'échantillon physique des produits reste nécessaire, ce qui maintient la contrainte de temps donnée par la distillation de l'effluent total. Ainsi, il a été proposé d'étudier la possibilité d'estimer les propriétés des distillats moyens à partir des informations spectrales de l'effluent total. Quatre propriétés du diesel (nombre de cétane, point d'écoulement, point de trouble, température limite de filtrabilité) et trois propriétés du kérosène (nombre de cétane, point d'éclair, point de fumée) ont été considérées. Cet objectif une fois atteint, l'optimisation du temps de réponse analytique était telle que l'estimation des propriétés pouvait être mise en œuvre dans le suivi et l'analyse des qualités produits issus du procédé en temps réel. En raison des avantages apportés par la spectroscopie proche infrarouge (PIR) pour répondre à la problématique posée, cette technique analytique a été le cœur du développement de la thèse.

Afin d'atteindre l'objectif de la thèse, la première question de recherche à laquelle il a fallu répondre fut : Est-il possible de prédire les propriétés des distillats moyens à partir des spectres PIR acquis sur l'effluent total produit lors de test sur unité pilote HCK ? Pour ce faire, 4 méthodes de régression (PLS, SVM, ANN, LWR) ont été évaluées dans la calibration de modèles prédictifs à partir de spectres PIR acquis sur des échantillons d'effluents totaux. Trois conclusions générales ont été tirées de ce travail. Premièrement, il est possible d'estimer les propriétés des distillats moyen à partir des spectres PIR acquis sur l'effluent total. La performance statistique des modèles développés (erreurs quadratiques moyennes de la validation croisée et de la prédiction) était proche ou même inférieure à la reproductibilité des méthodes de référence. Les modèles d'estimation du nombre de cétane du diesel et du kéroséne étaient les plus performants, tandis que les propriétés à froid du diesel étaient les plus exigeantes à modéliser. Deuxièmement, les performances obtenues à partir des différentes méthodes sont similaires, mais la régression PLS a l'avantage de produire un modèle interprétable et de présenter moins de risque de sur-apprentissage des modèles. Finalement, il convient de noter que malgré la performance acceptable des modèles développés, il était évident qu'il était possible d'améliorer la précision de l'estimation des propriétés étudiées. De plus, les performances des modèles sont affectées par l'évolution et la variabilité continues du procédé (charge, système catalytique, conditions opératoires), ce qui limite leur utilisation et leur fiabilité. Par conséquent, une compréhension plus approfondie des facteurs ayant un impact sur le comportement du procédé est nécessaire pour obtenir des estimations fiables.

Les performances limitées de certains modèles de prédiction ont conduit à la deuxième question de recherche : l'ajout d'informations supplémentaires et descriptives au spectre PIR améliore-t-elle les performances du modèle ? Pour répondre à cette question, une étape complémentaire de modélisation a

été réalisée, comprenant l'utilisation simultanée d'informations analytiques provenant de diverses sources (fusion de données) et la sélection de variables. Trois blocs de données ont été utilisés pour cette analyse : les spectres PIR et RMN acquis sur l'effluent total et les données du procédé.

Un premier étalonnage des modèles a été effectué en utilisant l'approche de fusion de données sans sélection de variables. Pour la modélisation de la fusion de données entre les deux blocs multivariés (spectres NIR et RMN), trois niveaux de fusion de données ont été évalués (bas, moyen et haut niveau). Pour le niveau bas de fusion, les méthodes de concaténation simple et SO-PLS ont été évaluées. Pour le niveau moyen de fusion, les scores de l'analyse PCA réalisés sur chaque bloc de données ont été utilisés comme caractéristique de fusion. À ce même niveau de fusion, le travail a été répété en utilisant les scores du modèle PLS calibrés à partir de chaque bloc de données pris séparément. Au niveau haut de fusion, la prédiction de la variable étudiée par les modèles PLS calibrés sur chaque bloc de données a été utilisée comme décision de fusion. Parmi les différents niveaux de fusion, le niveau moyen utilisant les scores des modèles PLS individuels était la meilleure stratégie pour améliorer les performances des modèles en fusionnant les informations de deux blocs spectroscopiques, mettant en avant la complémentarité des deux analyses. Pour la modélisation de la fusion des données entre les blocs multivariés et le bloc faiblement multivarié (données du procédé), les niveaux de fusion moyen et haut ont été évalués. Dans ce cas, le niveau haut de fusion a donné les meilleurs résultats. La fusion des données a amélioré les performances des modèles dans l'estimation de toutes les propriétés étudiées. Les modèles développés pour la prédiction des propriétés à froid de la coupe gazole ont présenté un gain important au niveau de leur performance.

Pour aller plus loin, une optimisation supplémentaire des modèles a été réalisée en utilisant la sélection des variables avant la calibration des modèles. Six méthodes de sélection de variables ont été évaluées sur les blocs PIR et RMN (VIP, SR, GA, iPLS, rPLS, CovSel). Par rapport aux méthodes évaluées, la méthode CovSel a sélectionné le nombre minimal de variables pour obtenir un modèle de prédiction dont la performance est comparable à celle du modèle utilisant toutes les variables. En ce qui concerne le bloc de données des variables du procédé, neuf méthodes ont été évaluées (VIP, SR, LASSO, GA, RFE, XGBoost_FS, SFS, SFFS, CovSel). Contrairement aux résultats montrés pour la sélection des variables dans les blocs multivariés, la performance de la méthode CovSel était limitée lorsqu'elle était évaluée sur le bloc des données du procédé. Pour ce bloc, la méthode qui a donné les meilleurs résultats a été le SFFS en sélectionnant les variables en mode inverse (backward selection). Après la sélection des variables dans chaque bloc de données, la modélisation de fusion des données a été répétée en utilisant le niveau haut de fusion. Les résultats ont fourni une estimation plus précise des propriétés étudiées, prédisant les propriétés à froid du diesel de tous les échantillons dans les limites de reproductibilité des méthodes de référence. Une nouvelle méthode de sélection de variables multi-blocs a également été évaluée (SO-CovSel). Cependant, la performance de cette méthode était limitée dans la fusion des données entre les blocs spectroscopiques et les données du procédé.

Bien que les performances des modèles aient été améliorées par la fusion des données et la sélection des variables, la fiabilité des modèles peut être affectée par des paramètres externes intervenant dans l'acquisition de l'information spectrale, en particulier lors des estimations en temps réel. Ainsi, une troisième question de recherche a été analysée : l'impact des paramètres externes sur la qualité des spectres PIR peut-elle être compensée/corrigée pour assurer une estimation en ligne fiable des propriétés ? Le travail développé pour répondre à cette question de recherche a été divisé en deux parties. La première partie correspondait à la définition des paramètres externes à étudier, et à l'acquisition des spectres aux conditions définies. Les paramètres identifiés ont été divisés en trois groupes : (i) modifications instrumentales, (ii) température de l'échantillon, et (iii) facteurs associés à l'acquisition en dynamique. Pour le premier groupe de paramètres externes évalués, deux types d'instruments (sonde et flowcell) avec deux trajets optiques différents (1mm et 2mm) ont été utilisés. En ce qui concerne la température de l'échantillon, 4 niveaux de température entre 60°C et 90°C avec un ΔT de 10°C ont été évalués. Pour le troisième groupe, un système de pompage en boucle fermée a été utilisé pour analyser la variation du débit et de la température de l'échantillon. La deuxième partie du travail correspondait à l'évaluation de différentes stratégies pour corriger l'impact des paramètres externes.

Quatre stratégies de correction ont été évaluées. La première stratégie consistait à générer une fonction de transfert qui corrigeait la déviation du spectre causée par les paramètres externes. La méthode PDS a été utilisée dans cette stratégie. La deuxième stratégie a consisté à développer un modèle de régression pour chaque paramètre externe étudié. En conséquence, 5 modèles PLS ont été développés pour l'estimation d'une même propriété. La troisième stratégie a consisté à développer un modèle de régression global en en intégrant les spectres acquis aux différentes conditions d'analyse dans l'ensemble d'étalonnage. Enfin, la quatrième approche s'est concentrée sur la modélisation robuste. Dans cette dernière approche, les méthodes d'orthogonalisation EPO et DOP ont été utilisées. La modélisation robuste intégrant les méthodes EPO et DOP a donné les meilleurs résultats pour corriger l'impact causé par les paramètres externes. Les modèles robustes obtenus ont permis d'avoir des estimations fiables des propriétés des distillats moyens quelles que soient les conditions d'acquisition, stables ou dynamiques, et même avec différents types d'instruments.

Pour valider le travail développé et les réponses aux questions de recherche proposées, deux cas d'étude ont été évalués. Le premier cas a évalué la performance des modèles développés pour prédire les propriétés des distillats moyens en utilisant des échantillons d'effluents totaux qui n'étaient pas inclus dans les ensembles de données initiaux de calibration et de validation des modèles. Certains de ces nouveaux échantillons ont été obtenus en traitant des charges et des systèmes catalytiques non pris en compte lors de la calibration du modèle. De plus, les spectres acquis sur ces échantillons l'ont été un an après l'acquisition des spectres utilisés pour la calibration des modèles. Les résultats de ce premier cas ont permis de valider les conclusions

faites précédemment. Plus précisément, il est possible de prédire les propriétés des distillats moyens à partir des informations spectroscopiques de l'effluent total. La précision de l'estimation peut être améliorée en utilisant des informations complémentaires aux spectres PIR. La fusion des données améliore la prédiction des propriétés, mais une amélioration plus importante est obtenue en utilisant la sélection des variables. Enfin, une modélisation robuste permet d'obtenir des estimations fiables des propriétés, quelles que soient les conditions d'acquisition.

Dans le deuxième cas évalué, un modèle robuste de prédiction de la densité de l'effluent total a été développé à partir des spectres acquis dans des conditions contrôlées du laboratoire, et utilisé pour suivre en ligne la stabilité opérationnelle d'une unité pilote HCK d'IFPEN. Les résultats ont montré la capacité du modèle à prédire correctement la densité de l'effluent total en temps réel. Au cours du suivi du test expérimental réalisé, il est devenu évident que le modèle évalue correctement l'impact des variations des variables opérationnelles sur la stabilité du procédé, ce qui permet d'identifier opportunément les déviations indésirables pendant le déroulement de l'essai. Dans le même cas d'étude, une estimation du nombre du cétane du diesel a été effectuée. Deux des prédictions faites ont été comparées à des valeurs mesurées en laboratoire. La déviation de l'estimation du nombre du cétane s'est avérée être inférieure à la reproductibilité de la méthode de référence. Ces résultats ont corroboré l'importance de la robustesse du modèle.

Les résultats obtenus dans ces travaux de recherche offrent une alternative fiable et robuste pour optimiser la recherche et le développement du procédé d'hydrocraquage. Comparé au schéma analytique conventionnel, le temps de réponse dans la caractérisation des distillats moyens en utilisant l'alternative étudiée pourrait être réduit de plusieurs semaines à quelques minutes. Par conséquent, les modèles développés peuvent être utilisés pour suivre la stabilité du procédé d'hydrocraquage en temps réel, favorisant ainsi une meilleure prise de décision. De plus, en ayant une connaissance en temps réel du comportement opérationnel du procédé et de son impact sur la qualité du produit, le chercheur peut décider quelles sont les analyses vraiment nécessaires, ce qui permet d'optimiser la recherche sur le procédé.

# Abstract (Français)

Cette thèse a étudié une alternative au schéma analytique classiquement appliqué pour caractériser les produits obtenus du procédé d'hydrocraquage. L'objectif principal de la recherche était donc de développer des modèles multivariés robustes à partir des informations spectroscopiques de l'effluent total d'hydrocraquage pour estimer les propriétés des distillats moyens (kérosène et gazole). Dans le cadre de ce projet, quatre propriétés du diesel (nombre de cétane, point d'écoulement, point de trouble, température limite de filtrabilité) et trois propriétés du kérosène (nombre de cétane, point d'éclair, point de fumée) ont été étudiés.

La faisabilité de l'estimation des propriétés des distillats moyens à partir des spectres proche infrarouge (PIR) acquis sur l'effluent total a d'abord été validée par des modèles de régression PLS. Les modèles développés ont présenté des erreurs proches ou même inférieures à la reproductibilité des méthodes analytiques de référence. Bien que les modèles PIR affichent des performances acceptables, ils mettent en évidence la nécessité d'une amélioration supplémentaire, notamment en ce qui concerne l'homoscédasticité et le coefficient de corrélation des modèles pour l'estimation des propriétés à froid du diesel. Ainsi, l'approche de modélisation par fusion de données a été appliquée pour améliorer les performances des modèles. Trois blocs de données ont été utilisés : les spectres PIR et RMN acquis sur l'effluent total et les données du procédé. En conséquence, l'estimation de toutes les propriétés a été améliorée. Les modèles de prédiction des propriétés à froid du diesel ont montré la plus grande amélioration. Une optimisation supplémentaire des performances du modèle a été obtenue en appliquant différentes méthodes de sélection des variables sur chaque bloc de données. En identifiant et en utilisant les variables les plus descriptives dans chaque bloc de données, il a été possible d'augmenter la précision de l'estimation des propriétés et de comprendre de manière exhaustive l'interaction et l'influence des variables indépendantes sur les propriétés étudiées.

Suite à la validation de la plausibilité de l'alternative étudiée, le problème du manque de robustesse des modèles a été abordé. Cette thèse a évalué l'impact des variations instrumentales, de la température de l'échantillon et des facteurs associés à l'acquisition dans des conditions dynamiques. Des modèles robustes ont été développés en utilisant les méthodes EPO (External Parameter Orthogonalization) et DOP (Dynamic Orthogonal Projection) pour corriger l'impact de ces paramètres. Les modèles robustes ont permis d'estimer de manière satisfaisante les propriétés des distillats moyens dans différents scénarios d'évaluation. Enfin, certains des modèles robustes développés ont été déployés pour suivre en temps réel la stabilité du procédé d'hydrocraquage, permettant une prise de décision opportune. Les résultats obtenus par ces travaux offrent une alternative fiable et robuste pour optimiser la recherche et le développement du procédé d'hydrocraquage. Comparé au schéma analytique traditionnel, le temps de réponse pour caractériser les distillats moyens en utilisant l'alternative étudiée pourrait être réduit de plusieurs semaines à quelques

minutes. De plus, en ayant une connaissance en temps réel du comportement opérationnel du procédé et de son impact sur la qualité du produit, le chercheur peut décider des analyses réellement nécessaires, ce qui permet d'optimiser la recherche sur le procédé.

# Abstract (English)

This thesis investigates an alternative to the analytical workflow normally followed when characterizing the products obtained from the hydrocracking process. Therefore, the main objective of the research was to develop robust multivariate models from spectroscopic information of the hydrocracking total effluent to estimate the properties of the middle distillates (kerosene and diesel). The work covered four properties of diesel (cetane number, pour point, cloud point, cold filter plugging point) and three properties of kerosene (cetane number, flash point, smoke point).

First, the feasibility of estimating middle distillate properties from near infrared (NIR) spectra acquired on the total effluent was validated through PLS regression models. The developed models presented errors close to or even lower than the reproducibility of the reference analytical methods. Although the NIR models had acceptable performance, they evidence the need for further improvement, particularly in the homoscedasticity and the correlation coefficient of the models for diesel cold flow properties estimation. Thus, the data fusion modelling approach was applied to improve the models' performance. Three data blocks were used: the NIR and NMR spectra acquired on the total effluent and the process variables. As a result, the estimation of all properties was enhanced. The models for predicting the diesel cold flow properties showed the greatest improvement. Further optimization in model performance was achieved by applying different variable selection methods to each data block. By identifying and using the most descriptive variables in each block of data, it was possible to increase the properties estimation accuracy and comprehensively understand the interaction and influence of the independent variables on the studied properties.

Following the plausibility validation of the investigated alternative, the lack of robustness of the models was addressed. This thesis evaluated the impact of instrumental changes, sample temperature, and factors associated with acquisition under dynamic conditions. Robust models were developed using the External Parameter Orthogonalization (EPO) and Dynamic Orthogonal Projection (DOP) methods for correcting the impact of these parameters. The robust models performed satisfactorily in estimating the middle distillate properties reliably under different evaluation scenarios. Finally, some of the developed robust models were deployed for monitoring in real-time the hydrocracking process stability, enabling timely decision-making. The results obtained from the research offer a reliable and robust alternative for optimizing the research and development of the hydrocracking process. Compared to the conventional analytical workflow, the response time in characterizing the middle distillates using the alternative investigated could be reduced from weeks to a few minutes. Furthermore, by having real-time knowledge of the process operation behavior and its impact on product quality, the researcher can decide on the actual analytics needed, leading to optimized process research.

# Table of Contents

# List of Figures

# List of Tables

# List of symbols and abbreviations

**μ:** Cinematic Viscosity

**AC:** Aromatic Carbon

**ANN:** Artificial Neural Networks

**AWLS-B:** Automatic Weighted Least Squares Baseline

**b:** model slope

**B-iPLS:** Backward interval PLS

**BPNN:** Back-Propagation Neural Networks

**BSC:** Bias and Slope Correction

**CFPP:** Cloud Filter Plugging Point

**CLS:** Classical Least Squares

**CN:** Cetane Number

**CovSel:** Covariance Selection

**CP:** Cloud Point

**DOP:** Dynamic Orthogonal Projection

**EMSC:** Extended MSC

**EPO:** External Parameter Orthogonalization

**FBP:** Final Boiling Point

**F-iPLS:** Forward interval PLS

**FP:** Flash Point

**G:** External parameter of influence

**g:** value external parameter of influence

**GA:** Genetic Algorithm

**GILS:** Genetic Inverse Least Squares

**H:** Hydrogen

**HCA:** Hierarchical Cluster Analysis

**HCK:** Hydrocracking

**HDT:** Hydrotreating

**IBP:** Initial Boiling Point

**IFPEN:** IFP Energies Nouvelles

**iPLS:** Interval PLS

**IR:** Infrared

**LASSO:** Least Absolute Shrinkage and Selection Operator

**LGO:** Light Gasoil

**LHSV:** Liquid Hourly Space Velocity

**LSS:** Loading Space Standardization

**LWR:** Locally Weighted Regression

**MC:** Mean Center

**MCR:** Multivariate Curve Resolution

**MCR-ALS:** Multivariate Curve Resolution - Alternating Least Squares

**MIR:** Mid-Infrared

**MLR:** Multiple Linear Regression

**MON:** Motor Octane Number

**MSC:** Multiplicative Scatter Correction

**N:** Nitrogen

**NC:** Naphthenic Carbon

**NCD:** Number of Calibration Data

**NIR:** Near-Infrared

**NMR:** Nuclear Magnetic Resonance

**NTD:** Number of Test Data

**NW-D:** Norris-Williams derivation

**OPLEC:** Optical Path Lenght Estimation and Correction

**P:** Pressure

**PC:** Paraffinic Carbon

**PCA:** Principal Components Analysis

**PCR:** Principal Component Regression

**PDS:** Piecewise Direct Standardization

**PLS:** Partial Least Squares

**PLS2-DA:** PLS2 Discriminant Analysis

**PNN:** Probabilistic Neural Networks

**PONA:** Paraffinics, Olefines, Naphthenes, Aromatics

**PP:** Pour Point

**PPR:** Projection Pursuit Regression

**PQN:** Probabilistic Quotient Normalization

**PV:** Process Variables

**$r^2$:** Pearson's squared correlation coefficient

**RF:** Random Forest

**RFE:** Recursive Feature Elimination

**RI:** Refractive Index

**RMSEC:** Root Mean Squared Error of Calibration

**RMSECV:** Root Mean Squared Error of Cross-Validation

**RMSEP:** Root Mean Squared Error Prediction

**RON:** Research Octane Number

**ROSA:** Response-Oriented Sequential Alternation

**rPLS:** recursive PLS

**RRM:** Reproducibility of the Reference Method

**S:** Sulphur

**SavGol:** Savitzky-Golay derivative

**SBFS:** Sequential Backward Floating Selection

**SBS:** Sequential Backward Selection

**SEP:** Standard Error of Prediction

**SFFS:** Sequential Forward Floating Selection

**SFS:** Sequential Forward Selection

**SimDis:** Simulated Distillation

**SNV:** Standard Normal Variate

**SO-CovSel:** Sequential Orthogonalized CovSel

**SO-PLS:** Sequential Orthogonalized PLS

**SP:** Smoke Point

**SR:** Selectivity Ratio

**SVM:** Support Vector Machine

**T:** Temperature

**VGO:** Vacuum Gasoil

**VIP:** Variable Importance in Projection

**VS:** Variable Selection

**VSN:** Variable Sorting for Normalization

**WNN:** Wavelet Neural Networks

**x:** independant variables

**XGBoost_FS:** eXtreme Gradient Boosting Feature Selection

**y:** dependant variables

**ŷ:** predicted variables

**ρ:** Density

# General Introduction

The increased demand for petroleum-based fuels has led to extensive and continuous research into the hydrocracking process. Most of the studies regarding this process are conducted to find the optimal operating conditions for maximizing the desired products while complying with the quality specifications required for their commercialization. Nonetheless, the extensive flexibility of the hydrocracking process makes its development, research, optimization, and innovation a high time- and cost-consuming labor.

The first research constraint is given by the time consumed in distilling the hydrocracking total effluent and the laboratory analysis conducted on the resulting cuts. The second constraint is the experimental cost involved in the laboratory analysis and the volume required for the total effluent distillation, resulting in the need to process a significant volume of feedstock in the pilot plant facilities.

The generation of mathematical models for feedstock and product properties estimation is a frequent practice in the oil & gas industry for process optimization and reduction of research costs. However, most of them are built based on other properties measured in the laboratory, thus maintaining the analysis response time constraint. Moreover, the models' performance is affected by the continuous evolution and variability of the process (feedstock, catalytic system, operating conditions), limiting their reliable use. Therefore, a more profound understanding of the factors impacting the process behavior is necessary for achieving reliable estimations.

Near-infrared spectroscopy (NIR) is a technique that presents several attractive characteristics to satisfy the described needs since it requires a low sample volume, has real-time responses, and contains extensive physicochemical information of the analyzed samples. Combining this analytical technique with chemometric methods has proven to be effective in developing predicting models for crude oil and its fractions properties estimation. This alternative for product characterization has helped reduce the time constraint given by the laboratory analysis. However, most chemometric models developed for fuel properties estimation employ the spectroscopic information acquired on the same stream being analyzed. Consequently, the distillation of the total effluent to obtain the physical cuts is still necessary, thus maintaining the time constraint given by this experimental task. Besides the limitation of the chemometric models regarding the need for the physical product sample to estimate the properties, it is important to stress that the low robustness of developed models leads to a relatively easy deterioration of their performance. Hence, the models need to be continuously adjusted and recalibrated.

Considering the given context, the thesis's main objective focused on optimizing the workflow employed for characterizing the fuels obtained from the hydrocracking process through reliable and accurate property estimation while avoiding the total effluent distillation. This objective addressed three research questions: (i) is it feasible to predict middle distillate properties from NIR spectra acquired on the total effluent obtained

from hydrocracking process reactors? (ii) including complementary and descriptive information to the NIR spectra improves the model performance?, and (iii) can external parameters' influence on the NIR spectra quality be compensated/corrected to ensure an online reliable properties estimation over time?. The work developed to answer these research questions led to obtaining robust chemometric models that estimate the properties of the middle distillates from the total effluent spectroscopic information. In this thesis, different regression models were developed for estimating four properties of diesel fuel and three properties of kerosene fuel. The obtained models have a statistical performance close to the reference methods used to measure the studied properties. The results drawn validated the feasibility of the characterization alternative investigated in this thesis, offering the potential to be deployed in real-time process monitoring applications.

The content of the manuscript is divided into six chapters. The first chapter gives the thesis context and the generalities and terminology used in the petroleum refining field. Next, details of the hydrocracking process and the motivation for its research, discussing the workflow followed traditionally in this labor, are outlined. The research problem and the alternatives reported in the literature to solve it are also discussed in this chapter. At last, the research questions addressed in this thesis are presented. The second chapter details the materials and methods used to answer the challenge. In chapter 3, the first research question is addressed by developing chemometrics models based on NIR spectra acquired on total effluent samples. This chapter compares different regression methods, defining the most suitable for the thesis purposes. The advantages and opportunities for improving the developed models' performance are also discussed. Chapter 4 shows the work done to improve the performance of the developed models through data fusion modelling, thus addressing the second research question. This chapter also discusses the advantages of using variable selection in property modelling. Chapter 5 discusses the issue of model robustness. This chapter answers the third research question by analyzing different approaches to correct the impact of external parameters on model performance. Chapter 6 presents the validation of the results and findings drawn in the previous chapters by implementing the models in two case studies. The first case study concerns evaluating the models' performance when the total effluent samples have been obtained under experimental conditions that were not considered during the models' calibration. The second case study evaluates the implementation of the models in the online monitoring of the hydrocracking process. Finally, conclusions and perspectives are presented.

It should be emphasized that the content of chapters 3 to 5 of the manuscript was based on the scientific publications produced in the thesis. Therefore, these chapters show the main results that helped to address the research questions, thus providing a coherent and practical thread to follow throughout these chapters. In addition, the scientific articles elaborated, some published, others submitted, and others in the submission

process, are attached as supplementary information in the appendices of the manuscript to give the reader a comprehensive detail of the results that led to the conclusions and perspectives of the thesis.

# CHAPTER I
# Introduction

# Chapter I. Introduction

This chapter has three main objectives. First, it seeks to introduce the reader to the domain of crude oil refining by presenting the general concepts and terminology used in this field. Then, it aims to give a general context of the relevance of petroleum product characterization in the research and optimization of the refining processes. Lastly, this chapter seeks to introduce the reader to the research problem addressed in the development of this thesis.

The first section of this chapter describes a general definition of petroleum, its main characteristics, and the most relevant refining processes for obtaining the final products. Next, this section presents the generalities of the hydrocracking process and its associated streams, emphasizing the total effluent and the middle distillates, the focus of the thesis research. In addition, the importance of characterizing the final product for decision-making in process optimization is highlighted. The second section of this chapter shows the studies reported in the literature on the different alternatives used to perform the product characterization task. Finally, a third section presents a general conclusion and the research questions that motivated the development of this thesis.

## 1. Research context

### 1.1 Petroleum and refining

Petroleum, also known as crude oil, is a complex mixture of hydrocarbons and heteroatomic compounds such as sulfur, oxygen, and nitrogen. The concentration of each component can vary according to the oil's origin. Hydrocarbon molecules contribute the highest percentage in this mixture (90-99 %wt)[1]. These compounds can be classified into four main groups according to their chemical nature: paraffins, olefins, naphthenes and aromatics (PONA). The fraction of each group in the oil is also variable[1]. The other compounds are found in smaller proportions in the crude oil and, being undesirable compounds, are considered contaminants. Sulfur (0.01-6 %wt) and nitrogen (0.05-0.5 wt%) compounds have a deleterious effect on the catalytic refining processes (they poison the catalyst) and impact the product quality obtained. The oxygenated compounds (0.1-0.5%wt) and other contaminants such as metals, mainly nickel and vanadium (0.005 - 0.15% wt)[1], are also sought to be removed. The crude oil can be classified based on two physicochemical properties: API gravity and sulfur content. API gravity determines whether the crude oil is light (°API > 35 ) or heavy (°API < 18 ), and the sulfur establishes whether the petroleum is sweet (< 0.5 %wt) or sour (> 1%)[2].

Crude oil is currently the primary energy source and feedstock for generating different products, from gasoline, kerosene, and diesel to plastics and textiles for example. For obtaining those products, crude oil requires different treatment and transformation processes, known as refining processes. Figure 1 shows a

general diagram of a refinery where it is observed that the crude oil is initially sent to the process of atmospheric distillation from which gas, the fuels already mentioned, atmospheric gas oil, and residue are obtained. The naphtha is hydrotreated to remove sulfur compounds. Next, this hydrotreated naphtha is sent to other processes such as isomerization and catalytic reforming to obtain gasoline that finally meets the required quality for commercialization. Kerosene or jet fuel obtained is generally sent to a complementary process for mercaptans removal. Diesel oil is also hydrotreated to remove sulfur compounds and contaminants.

The residue is sent to another distillation process operated under vacuum conditions, thus obtaining vacuum gas oil (VGO) and its respective residue. Conversion processes, whether thermal such as Visbreaking and Delayed Coking, or catalytic such as Fluid Catalytic Cracking (FCC) and Hydrocracking (HCK), are then used to convert these cuts by cracking the long carbon chains into smaller ones. These lighter cuts are then hydrotreated to fulfill market requirements.

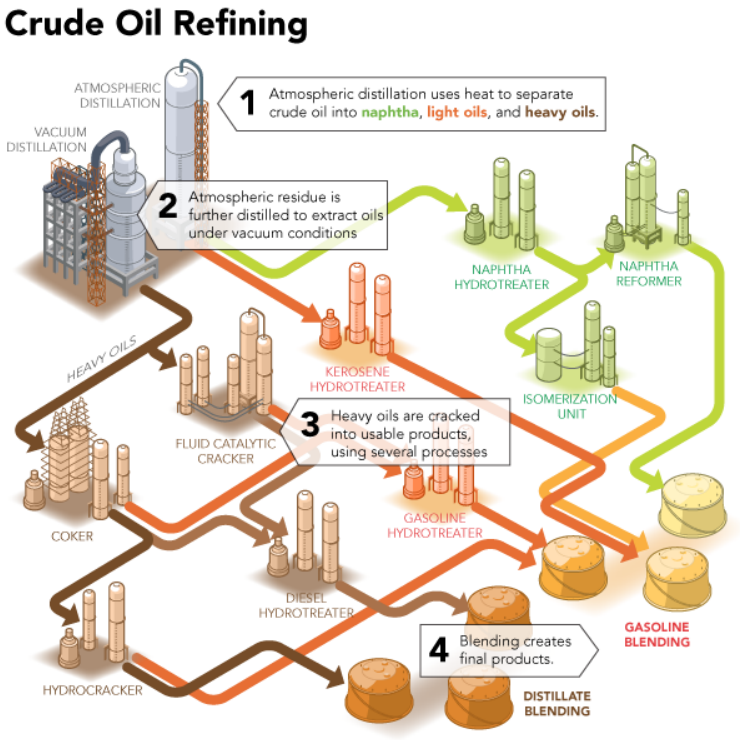



*Figure 1. General flow diagram of a refinery*
*'https://blog.gltproducts.com/blog/what-goes-on-at-an-oil-refinery'*

The shift in consumption from gasoline to diesel has led over the last 20 years to a strong worldwide increase in demand for middle distillates (kerosene and diesel)[3]. At the same time, the increasing heavy crude oil production[4] has resulted in low-quality feedstocks being processed. The outlined issues and the constant demand for high-quality products have led refineries to require flexible refining processes that maximize the production of middle distillates from heavy feedstocks while ensuring their quality for compliance with environmental and commercial legislations[5, 6]. Given its extensive flexibility, the hydrocracking (HCK) process

is essential in addressing the needs described [7].

## 1.2 Hydrocracking process and research motivation

Hydrocracking is a catalytic process that aims to convert heavy feedstocks (long-chain molecules with high-boiling points) into lighter and high-quality products (lower-boiling points and short-chain molecules) through cracking of the carbon-carbon bonds and the hydrogenation of the shorter chains generated from this cracking[8]. The HCK process is commonly used for upgrading heavier fractions obtained from crude oil distillation, including the residue. It is also used to upgrade products from other processes, such as coker gasoil, deasphalted oil, FCC cycle oils, and tower bottoms. This process has emerged as the primary diesel producer in many refinery configurations addressing the need for maximizing the production of middle distillates from heavy feedstocks. Unlike the FCC process, HCK can effectively yield ultra-low sulfur diesel (ULSD), whereas middle-distillate range FCC products regularly require additional treatment to meet product environmental and commercial specifications[9]. As evidenced, HCK is a refining process with extensive flexibility in processing heavy feedstocks to obtain various high-quality fuel products[10]. This flexibility is linked to the operating conditions employed in the process, primarily the formulation of the catalytic system[11].

The hydrocracking process may consist of up to 6 configurations: one or two stages with no, partial, or total recycling[12]. Depending on the operating conditions (pressure and temperature), it can be categorized as mild or high-pressure hydrocracking. Despite the high investment cost (CAPEX) and its significant operating cost (OPEX), refiners express a high interest in high-pressure HCK because it has more flexibility in the process, and the middle distillates obtained have better quality (cetane number and cold flow properties for diesel). Table 1 summarizes the operating conditions used in the different schemes of this process. The most common and analyzed scheme in this thesis is depicted in Figure 2. In broad terms, the process has a first hydrotreatment (HDT) stage that removes heteroatoms, saturates the olefins, and partially hydrogenates the aromatics. Subsequently, the hydrotreated feedstock is sent to a reactor where, in the presence of a specific catalyst, the HCK reactions occur[13]. Finally, a lighter liquid product known as total effluent is obtained from the reaction section and distilled to obtain the desired products. In Figure 2 are shown the streams involved in the process: hydrogen, feedstock (mostly VGO), and the hydrocracked total effluent obtained from the HDT and HCK reactors.

*Table 1. HCK process operating conditions[14]*

| Hydrocracking unit type | Typical conversion, % | Total Pressure, Bar | Hydrogen partial pressure, Bar | Reactor temperature, °C (°F) |
|---|---|---|---|---|
| Mild | 20-40 | 60-100 | 20-55 | 350-440 (662-824) |
| Moderate/Medium-pressure | 40-70 | 100-110 | 50-95 | 340-435 (644-815) |
| Conventional/High-pressure | 50-100 | 110-200 | 95-140 | 350-450 (662-842) |
| Resid hydrocracking | 65-100 | 97-340 | 73-255 | 385-490 (725-914) |

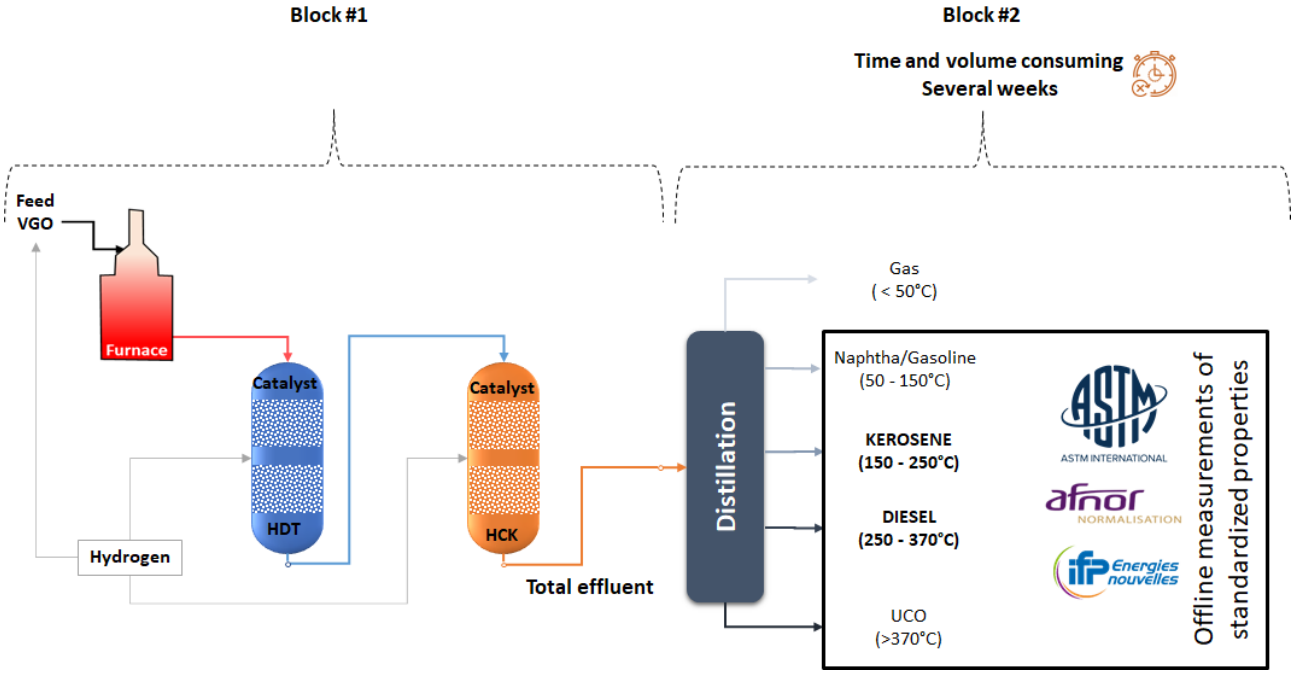



*Figure 2. Hydrocracking process global scheme*

As discussed previously, the HCK process is an extensively implemented refining process nowadays. Hence, it is the subject of ongoing research. In their study, "Hydrocracking: A Perspective towards Digitalization", Iplik et al.,[15] showed the exponential increase in the number of scientific publications related to the hydrocracking process in the last two decades. Approximately 89% of the reported studies are related to process optimization focusing on three research lines: applied chemistry in catalyst development (30%), evaluation of sequencing system configurations and operational conditions (58%), and automation and control systems (1%)[15]. Considering these three research fronts, the variability of the HCK process to be investigated can be substantial, making the development and innovation of this process a cost- and time-consuming task.

The HCK process variability is commonly evaluated by conducting different experimental designs in pilot plants and laboratory facilities under various controlled conditions. Generally, the experimentation is divided into the catalytic reaction and distillation steps. The first step, summarized in block #1 of Figure 2, concerns the heating and conversion of the feedstock-hydrogen mixture into a lighter product (total effluent) through the HDT and HCK reactions. The impact of process operating conditions such as pressure (P), temperature (T), the residence time given by the liquid hourly space velocity (LHSV), and the catalytic system are screened during this step. This task is conducted in an uninterrupted sequence over the entire study's length. The second step, summarized in block #2 of Figure 2, involves the distillation of the total effluent to obtain the final products, including the middle distillates, for subsequent characterization. Finally, the resulting analytical information is gathered and analyzed for process monitoring and evaluation.

In contrast to the catalytic reaction step, the characterization of the product is performed on a discontinuous

time basis. Firstly, the laboratory analyses are conducted offline and are conditioned to the different laboratories' response times. Moreover, the physical product samples are required to perform the laboratory analysis. The samples are obtained from the total effluent distillation conducted in a non-continuous sequence.

Depending on the study purpose, the experimental scheme described requires processing considerable amounts of feedstock and catalyst to produce a sufficient volume of total effluent for its physical fractionation. To reduce the consumption of these supplies in the HCK process research, High Throughput Experimental units (HTE)[16] are employed. They use small sample volumes of feedstock and catalyst and allow simultaneous process analysis (parallel studies/tests), helping the cost-benefit ratio of research projects. However, when using HTE units, a sufficient volume of total effluent is often not produced, limiting the detailed process analysis regarding product characterization. In addition, it should be noted that some HTE unit configurations allow producing the minimum required volume of total effluent for distillation; nevertheless, the volume of final cuts obtained is sometimes insufficient to perform a complete laboratory analysis. A common practice in HCK process research for optimizing experimental tests performed in pilot units is developing kinetic models and simulators. However, these models and simulators must be periodically fed with experimental test data to avoid deterioration and obsolescence.

The characterization of the products obtained from the HCK process is a crucial but time- and cost-consuming task in its research. This labor can be accomplished by implementing different approaches. The most commonly employed approach is using standardized norms and methods. An alternative to optimize costs and response time is substituting these standardized analyses with multivariate analytical techniques coupled with statistical processing. The following section provides an overview of the latter alternative and its application in the oil & gas sector.

# 2. Product characterization approaches

## 2.1 Standard methods approach

Whether for research or quality validation, the characterization of petroleum cuts must be conducted in a reliable, repeatable, and reproducible manner. Therefore, the characterization must comply with certain standardized guidance that can be applied at the international or world level. The most widely used analytical methods in the oil industry are the American Society of Testing Materials (ASTM) and the International Organization for Standardization (ISO) methods; yet these methods can be adapted to specific needs according to the legislation of each region. Table 2 presents the most representative laboratory analyses used to determine the quality of the petroleum products involved in the development of this thesis.

The standard methodologies defined by ASTM and ISO require trained operators, an infrastructure that

guarantees the measuring conditions, a system for assuring results and specialized equipment to carry out the different tests according to the standard. In addition, the time required to obtain the results can vary between half an hour and eight hours, not considering the time required to distill the total effluent. Hence, the characterization of the HCK process streams according to the outlined standard norms restricts the opportune analysis and monitoring of the process. Therefore, a fast and efficient alternative for the crude oil cuts characterization is of great interest.

*Table 2. Standard analytical methods for quality determination of HCK process streams (feedstock, total effluent, middle distillates)*

| Property | Stream | Method | Reproducibility limits | |
|---|---|---|---|---|
| Viscosity @ 70°C & 100°C (μ) - cst | Feedsctock | ASTM D445-97[17] | ± 0.0082 (μ_measured + 1) | |
| Sulphur (S) - wt% | | ISO 20846[18] | ± 0.112(S_measured)+1.12 | |
| Nitrogen (N) - mg/kg ppm | | ASTM D5291[19] | ± 0.4456 | |
| Hydrogen (%H) - w/w% | | | ± 0.2314(H_measured$^{0.5}$) | |
| Aromatic Carbon (AC) - wt% | | ASTM D3238-95[20] | ± 1.7 | |
| Paraffinic Carbon (PC) - wt% | | | ± 3.4 | |
| Naphtenic Carbon (NC) - wt% | | | ± 3.6 | |
| Density (ρ) - gr/ml | Feedstock Total effluent Middle distillates | ASTM D1218 - 12[21] | ±0.0005 | |
| Refractive Index (RI) @20°C | | | ±0.0002 | |
| Refractive Index (RI) @70°C | | ASTM D1747[22] | ±0.0006 | |
| Simulated distillation (SimDis) °C | Feedstock | ASTM D7213-15[23] | | |
| | Total effluent Middle distillates | ASTM D2887-19ae2[24] | | |
| Cetane Number (CN) | Middle distillates | ASTM D613-01[25] | **Average CN** | **Limits** |
| | | | 40 | ± 2.8 |
| | | | 44 | ± 3.3 |
| | | | 48 | ± 3.8 |
| | | | 52 | ± 4.3 |
| | | | 56 | ± 4.8 |
| Flash Point (FP) - °C | Kerosene | ASTM D93-18[26] | ± 0.071*(FP_measured) | |
| Smoke Point (SP) - mm | | ASTM D1322-12[27] | ± 0.001651*(SP_measured+30) | |
| Cloud Point (CP) - °C | Diesel | NF EN 23015[28] | ± 4 | |
| Pour Point (PP) - °C | | ASTM D5949[29] | ± 6 | |
| Cold Filter Pluging Point (CFPP) - °C | | NF EN 116[30] | ± 3-0.06*(FLT measured) | |

## 2.2 Alternative approach

In the last decades, combining analytical analysis and chemometric methods has drastically increased to assess fuels, from crude oils to refined cuts such as gasoline[31], diesel[32, 33] and biodiesel[34, 35], or lubricants[36]. On the one hand, the main advantage of applying multivariate calibration methods to analytical techniques is both cost- and time-saving. On the other hand, the sample volume required is quite low compared to some standardized methods used to characterize fuels.

Among the known existing analytical methods, the vibrational spectroscopy is the most adequate to comply with the described optimization needs, highlighting the infrared spectroscopy[37] (IR), either near (NIR) or mid (MIR). NIR is one of the most relevant and attractive techniques for developing predictive models as it is a non-destructive, non-invasive method, requires a small sample volume, requires minimal sample preparation, and is suitable for applications where real-time measurement is required[38, 39]. Furthermore, this analytical technique is highly flexible and used in research and industrial applications, from food quality analysis to chemical process control[40].

Another analytical technique for property estimation from regression models is the Nuclear Magnetic Resonance (NMR). The most important applications in organic chemistry are NMR spectrometry of proton ($^{1}H$) and carbon-13 ($^{13}C$). It is a powerful and versatile method that can be applied to solid and liquid materials and quickly analyze samples requiring minimal preparation[41]. The fundamental application of NMR spectroscopy is the structural determination of either organic, organometallic, or biological molecules. Compared to infrared spectroscopy, the NMR spectrum contains more detailed information on the molecular interactions and bonds present in the sample.

The outlined analytical methods have characteristics that make them attractive for the purposes aforementioned. Table 3 illustrates and compares the main features of each of them.

*Table 3. Main features of NIR and NMR techniques*

| | **NIR** | **NMR** |
|---|---|---|
| Molecular sample description | ✗ | ✓ |
| Prediction of samples' physicochemical properties | ✓ | ✓ |
| Low sample volume | ✓ | ✓ |
| Minimal ~No sample preparation | ✓ | ✗ |
| Analysis time <2h | ✓ | ✓ |
| Real-time application | ✓ | ✗ |

The growing need to take advantage of the large volume of information generated by the analytical techniques described has made chemometrics, which is defined as a "chemical discipline that uses mathematical, statistical and logical methods to extract valuable information from experimental data to optimize processes and/or products"[42], very popular in the development of prediction models as an alternative for property characterization. To obtain reliable prediction models, it is fundamental to ensure the quality and suitability of the database utilized. This includes, but is not limited to, preprocessing of information, detecting anomalous data, and further statistical analysis of the database. In addition to ensuring an appropriate and reliable database, it is crucial to adequately select the chemometric methods to be used when calibrating the models. The methods used should optimally explain the relationship between the information extracted from the analytical technique and the property investigated. As mentioned at the beginning of this section, several studies have focused on applying chemometric methods to analytical techniques for petroleum product characterization. In the next pages, an analysis of these studies is given to

present the most relevant progress made on this subject and outline the research problem to be addressed in this line of work. It is important to point out that the methods employed in the studies analyzed are not discussed in this section. Instead, a general description of them and their respective references are reported in appendix 8 as complementary information for the reader.

A recent review from Moro et al.[43] points out the growing use of infrared spectroscopy and NMR to predict crude oil properties using chemometrics methods. They show in their study the compilation of 35 studies completed between 1998 and 2020. This compendium of studies focuses on estimating 24 crude oil properties, being the API gravity, nitrogen, and sulfur content the most investigated. NIR spectroscopy was used in 23% of these studies, 30% used MIR, and the remaining 47% used NMR.

According to the literature analyzed in this thesis, there is no existing equivalent review for other petroleum cuts. However, several interesting studies can be found showing the interest in using IR and/or NMR and chemometrics to rapidly obtain properties of fuels with statistical performance close to the reference methods. Following the consolidation and analysis scheme made by Moro et al.[43], Table 4 shows a non-exhaustive compilation of these works.

*Table 4. Chemometrics studies applied for fuel properties estimation*

| Year | Spectral input data | Cut analyzed | Property estimated | Number of samples | Regression Method | Preprocessing | Reference |
|---|---|---|---|---|---|---|---|
| 1990 | NIR | Gasoline | RON<br>MON<br>PONA (wt%) | 28 | PLS | Baseline correction | Parisi[44] |
| 1995 | MIR | Kerosene | Density (g/mL)<br>Freezing point (°C)<br>Flash point (°C)<br>Aromatic content (wt%)<br>Initial Boiling Point (IBP) (°C)<br>Final Boiling Point (FBP) (°C)<br>Viscosity (cSt) | 29 | PLS | mean centering | Garrigues[45] |
| 1997 | MIR | Kerosene | Density (g/mL)<br>Freezing Point (°C)<br>Flash Point (°C)<br>Aromatics Content (v/v%)<br>Initial Boiling Point (IBP) (°C)<br>Final Boiling Point (FBP (°C) | 29 | PLS, PCR, MLR | mean centering | Andrade[46] |
| 1998 | MIR<br>NMR | Base Oil | Viscosity Index<br>Pour Point (°C)<br>Carbon Type | 60 | PLS | mean centering | Sastry[47] |
| 1999 | NIR | Diesel & Kerosene | Cetane Number<br>Refractive Index @20°C<br>Density (g/ml)<br>Hydrogen content (wt%)<br>Aromatic Carbon (wt%)<br>Aromatics Content (wt%) | 90 | PLS | mean centering | Zanier-Szydlowski[48] |
| 1999 | NIR & MIR | Kerosene | Distillation Curve | 50 | PLS | mean centering | Chung[49] |

*Table 4. Chemometrics studies applied for fuel properties estimation continuation*

| Year | Spectral input data | Cut analyzed | Property estimated | Number of samples | Regression Method | Preprocessing | Reference |
|---|---|---|---|---|---|---|---|
| 2000 | NIR | LSR | Model developed to classify the petroleum cuts | 57 | PCA & Bayesian classifier | SNV | Kim[50] |
| | | Naphtha | | 61 | | | |
| | | Kerosene | | 61 | | | |
| | | Diesel | | 64 | | | |
| | | LGO | | 64 | | | |
| 2001 | NMR | Gasoline | RON | >300 | ANN | mean centering | Meusinger[51] |
| | | | MON | | | | |
| 2003 | MIR | Kerosene | Flash point (°C) | 100 | PLS | mean centering | Gômez-Carracedo[52] |
| | | | Freezing point (°C) | | | | |
| | | | Initial Boiling Point (IBP) (°C) | | | | |
| | | | 10% of distilled sample (°C) | | | | |
| | | | 90% of distilled sample (°C) | | | | |
| | | | Final Boiling Point (FBP) (°C) | | | | |
| | | | Aromatics (wt%) | | | | |
| | | | Viscosity (cSt) | | | | |
| 2003 | NMR | Diesel | Cetane Number | 60 | PCA-ANN | mean centering | Basu[53] |
| 2004 | MIR | Lubricating Oil | Contaminants (gasoline, ethylene glycol, water) content (wt%) | 78 | iPLS | MSC | Borin[54] |
| 2006 | MIR | Kerosene & Diesel | Aromatics content (wt%) | 59 | PLS | mean center | Baldrich[55] |
| | | | Sulfur content (ppm) | | | | |
| | | | 50% of distilled sample (°C) | | | | |
| | | | Final Boiling Point (FBP) (°C) | | | | |
| 2007 | NIR | Gasoline | Density (g/mL) | 106 | PLS PCR ANN MLR | normalization autoscaling | Balabin[56] |
| | | | Initial Boiling Point (IBP) (°C) | | | | |
| | | | 10% of distilled sample (°C) | | | | |
| | | | 50% of distilled sample (°C) | | | | |
| | | | 90% of distilled sample (°C) | | | | |
| | | | Final Boiling Point (FBP) (°C) | | | | |
| 2008 | NIR | Gasoline | Density (g/mL) | 227 | WNN | No preprocessing | Balabin[57] |
| | | | Benzene content (ppm) | | | | |
| | | | Ethanol content (ppm) | | | | |
| 2008 | NIR | Biodiesel | Iodine value | 311 | PLS | Savitzky-Golay [9,3,1-2] | Baptista[58] |
| | | | CFPP (°C) | 71 | | | |
| | | | Viscosity (cst) | 144 | | | |
| | | | Density (g/mL) | 91 | | | |
| 2008 | NIR | Diesel | Cetane Number | 245 | GILS | 1st derivative | Özdemir[59] |
| | | | Boiling Point (°C) | 246 | | | |
| | | | Freezing point (°C) | 251 | | | |
| | | | Aromatics (wt%) | 256 | | | |
| | | | Viscosity (cSt) | 252 | | | |
| | | | Density (g/mL) | 263 | | | |
| 2008 | NIR | Gasoline | RON | 156 | PLS (calibration transfer) | SNV MSC | Pereira[60] |
| | | | Naphthenes content (wt%) | | | | |
| 2009 | NIR | Gasoline | RON | 67 | PLS (calibration transfer) | SNV | Amat-Tosello[61] |
| | | | MON | | | | |
| 2010 | MIR | Motor Oil | Viscosity Index | 30 | PLS | Savitzky-Golay [10,1,0] | Al-Ghouti[62] |
| | | | Base number | | | | |
| 2010 | NIR | Motor Oil | Model developed to classify motor oil by base stock | 225 | PNN, SVM | mean centering | Balabin[63] |

*Table 4. Chemometrics studies applied for fuel properties estimation continuation*

| Year | Spectral input data | Cut analyzed | Property estimated | Number of samples | Regression Method | Preprocessing | Reference |
|---|---|---|---|---|---|---|---|
| 2010 | NIR | Gasoline | RON | 384 | LPC & MLR | Normalization | Kardamakis[64] |
| 2010 | NIR | Diesel | Density (g/mL) | 161 | PLS | Savitzky-Golay [11,2,1] | Fátima[65] |
| | | | Sulfur content (ppm) | | | | |
| | | | Distillation temperatures °C | | | | |
| 2010 | NIR | FCC feedstocks | Sulfur (wt%) | 89 | PLS | mean center | Baldrich[66] |
| | | | Density (kg/L) | | | | |
| | | | Basic Nitrogen (wt%) | | | | |
| | | | Microcarbon Residue (wt%) | | | | |
| | | | Nickel (ppm) | | | | |
| | | | Vanadium (ppm) | | | | |
| 2011 | NIR | Jet Fuel | API gravity | 70 | PLS (calibration transfer) | SNV | Cooper[67] |
| | | | Aromatics content (%wt) | | | | |
| | | | Cetane index | | | | |
| | | | Density (g/mL) | | | | |
| | | | 10% of distilled sample (°C) | | | | |
| | | | 20% of distilled sample (°C) | | | | |
| | | | 50% of distilled sample (°C) | | | | |
| | | | 90% of distilled sample (°C) | | | | |
| | | | Flash point (°C) | | | | |
| | | | Hydrogen content (%wt) | | | | |
| | | | Saturates content (%wt) | | | | |
| | | | Viscosity (cSt) | | | | |
| 2011 | NIR | Biodiesel | CFPP (°C) | 101 | PLS | Savitzky-Golay [9,3,2] | Balabin[68] |
| | | | Iodine value | | | | |
| 2011 | NIR | Biodiesel | Density (g/mL) | 124 | ANN | Savitzky-Golay OSC | Balabin[69] |
| | | | Viscosity (cSt) | | | | |
| | | | Methanol Content (ppm) | | | | |
| | | | Water content (ppm) | | | | |
| 2012 | MIR | Lubricating Oil | Density (g/mL) | 100 | PLS | mean centering | Marinovic[70] |
| | | | Viscosity (cSt) | | | | |
| | | | Pour Point (°C) | | | | |
| 2012 | MIR | Motor Oil | Adulteration grade | 60 | PLS2-DA | SNV | Bassbas[71] |
| 2012 | NIR | Diesel | Cetane Number | 245 | PLS | mean centering | Yan-Kun[72] |
| 2012 | MIR | Diesel | Cetane Number | 93 | PLS | mean centering | Marinovic[73] |
| | | | Cetane Index | | | | |
| | | | Density (g/mL) | | | | |
| | | | Viscosity (cSt) | | | | |
| | | | 10% of distilled sample (°C) | | | | |
| | | | 50% of distilled sample (°C) | | | | |
| | | | 90% of distilled sample (°C) | | | | |
| | | | Aromatics content (wt%) | | | | |
| 2012 | NIR | JetFuel | Flash Point (°C) | 60 | PLS | First derivative | Xu[74] |
| | | | Freezing Point (°C) | | | | |
| | | | 10% of distilled sample (°C) | | | | |
| | | | 50% of distilled sample (°C) | | | | |
| | | | 90% of distilled sample (°C) | | | | |
| 2013 | Vis-NIR | Lubricating oil | Insoluble content (wt%) | 70 | PLS (Variable selection) | Savitzky-Golay [5,1,2] MSC | Villar[75] |

*Table 4. Chemometrics studies applied for fuel properties estimation continuation*

| Year | Spectral input data | Cut analyzed | Property estimated | Number of samples | Regression Method | Preprocessing | Reference |
|---|---|---|---|---|---|---|---|
| 2014 | NMR | Diesel | Cetane Number | 60 | PLS | mean centering | Souza[76] |
| 2015 | NIR | Gasoline | Density (g/ml) | 466 | Adaptive Algorithm ORL-PLS (Author's creation) | 1st order derivative | He[77] |
| | | | Freezing Point (°C) | | | | |
| | | | Aromatics Content (wt%) | | | | |
| | | | Viscosity (cSt) | | | | |
| | | | Cetane Number | | | | |
| 2015 | NMR | Diesel | Cetane Index | 60 | PLS | - | Santos[78] |
| | | | Density (g/mL) | | | | |
| | | | Flash Point (°C) | | | | |
| | | | 50% of distilled sample (°C) | | | | |
| 2015 | NIR | Diesel | Freezing point (°C) | 441 | LS-SVM/PLS | Savitzky-Golay MSC | Feng[79] |
| | | | Density (g/mL) | | | | |
| | | | Viscosity (cSt) | | | | |
| | | | Boiling Point (°C) | | | | |
| | | | Cetane Number | | | | |
| | | | Aromatics (wt%) | | | | |
| 2016 | NMR | Diesel | Cetane Number | 125 | MLR | - | Abdul Jameel[80] |
| 2016 | NIR | Shale oil | Density (g/mL) | 300 | PLS (calibration transfer) | Cubic spline | Baird[81] |
| 2016 | NIR | Diesel | Density (g/mL) | 166 | PLS | Savitzky-Golay [11,3,1] | Brouillette[82] |
| | | | Cetane Index | 141 | | | |
| | | | Viscosity (cSt) | 134 | | | |
| | | | Aromatics (wt%) | 35 | | | |
| | | | Cloud Point (°C) | 111 | | | |
| | | | Flash Point (°C) | 107 | | | |
| | | | Pour Point (°C) | 95 | | | |
| | | Kerosene | Density (g/mL) | 89 | | | |
| | | | Aromatics (wt%) | 50 | | | |
| | | | Flash point (°C) | 92 | | | |
| | | | Pour Point (°C) | 44 | | | |
| | | | Freezing point (°C) | 86 | | | |
| 2017 | NIR | Gasoline + Ethanol | Ethanol Content (v/v%) | 23 | MCR-ALS | Savitzky-Golay [9,1,2] normalization | Oliveira[83] |
| 2017 | NIR & MIR | Bio-Diesel | Distillation Curve | 16 | PLS | Savitzky-Golay [7,1,0] | Câmara[84] |
| | | | Viscosity (cSt) | | | | |
| | | | Flash point (°C) | | | | |
| | | | Water content (ppm) | | | | |
| 2017 | MIR | Motor Oil | Total Acid Number (TAN) | 80 | PLS, SVM, RF, PPR | normalization | Leal de Rivas[85] |
| 2017 | MIR NIR | Diesel biodiesel | Fatty methyl esters | 50 | Data fusion | Savitzky-Golay SNV | Luna[86] |
| 2017 | MIR | Crude oil | API gravity | 96 | PLS (calibration transfer) | Mean center | Rodrigues[87] |
| 2017 | NIR | Diesel | Cetane Number | 381 | LS-SVM | Savitzky-Golay [15,1,1] | Zhan[88] |

*Table 4. Chemometrics studies applied for fuel properties estimation continuation*

| Year | Spectral input data | Cut analyzed | Property estimated | Number of samples | Regression Method | Preprocessing | Reference |
|---|---|---|---|---|---|---|---|
| 2017 | NIR | Diesel | Density (g/mL) | 278 | PLS | Savitzky-Golay [11,2,0] | Palou[89] |
| | | | Cetane Index | | | | |
| | | | FAME | | | | |
| | | | Cloud Point (°C) | | | | |
| | | | 95% of distilled sample (°C) | | | | |
| | | | Flash Point (°C) | | | | |
| | | | Sulfur content (ppm) | | | | |
| 2017 | NIR | Gasoline | 10% of distilled sample (°C) | 103 | PLS (calibration transfer) | SNV MSC | Da Silva[90] |
| | | | 50% of distilled sample (°C) | | | | |
| | | | 90% of distilled sample (°C) | | | | |
| | | | Final Boiling Point (FBP) (°C) | | | | |
| | | | Density (g/mL) | | | | |
| 2018 | MIR | Diesel | Density (g/mL) | 409 | PLS (Variable selection) | OSC | Nespeca[32] |
| | | | Flash Point (°C) | | | | |
| | | | Total Sulfur (ppm) | | | | |
| | | | Distillation curve (°C, v/v%) | | | | |
| | | | Cetane Index | | | | |
| 2018 | NMR | Diesel | Cloud Point (°C) | 40 | ANN, Kriging | normalization | da Costa Soares[91] |
| | | | CFPP (°C) | | | | |
| | | | Viscosity Index | | | | |
| 2018 | NMR | Kerosene | Smoke Point (°C) | 197 | PLS | Savitzky-Golay normalization | Lacoue-Nègre[92] |
| | | | Cetane | 225 | | | |
| | | | Hydrogen content (wt%) | 204 | | | |
| | | | Mono-Aromatics (wt%) | 202 | | | |
| | | | Di-Aromatics (wt%) | 71 | | | |
| | | | Total-Aromatics (wt%) | 196 | | | |
| | | Diesel | Cloud Point (°C) | 276 | | | |
| | | | Pour point (°C) | 203 | | | |
| | | | CFPP (°C) | 254 | | | |
| | | | Cetane | 290 | | | |
| | | | Hydrogen content (wt%) | 279 | | | |
| | | | Mono-Aromatics (wt%) | 221 | | | |
| | | | Di-Aromatics (wt%) | 221 | | | |
| | | | Total-Aromatics (wt%) | 217 | | | |
| | | 370°C$^{+}$ | Hydrogen Content (wt%) | 281 | | | |
| | | | Carbon Content (wt%) | 269 | | | |
| | | | Pour Point (°C) | 139 | | | |
| | | | Viscosity Index | 344 | | | |
| | | Lubricant Oil | Pour Point + Viscosity Index | 203 | | | |
| 2019 | NMR | Diesel | Viscosity | 40 | PLS | mean centering | Constantino[93] |
| | | | Density | | | | |
| | | | Refractive Index | | | | |
| 2020 | NIR | Diesel | Cetane Number | 50 | PLS | Baseline correction | Barra[94] |
| 2020 | NIR | Diesel | Cetane Number | 784 | Regression tree (Variable selection) | mean centering | Shukla[95] |
| | | | Boiling point (°C) | | | | |
| | | | Freezing point (°C) | | | | |
| | | | Aromatics content (wt%) | | | | |
| | | | Viscosity (cSt) | | | | |
| | | | Density (g/mL) | | | | |

*Table 4. Chemometrics studies applied for fuel properties estimation continuation*

| Year | Spectral input data | Cut analyzed | Property estimated | Number of samples | Regression Method | Preprocessing | Reference |
|---|---|---|---|---|---|---|---|
| 2021 | NIR | Diesel | Boiling Point (°C) | 237 | PLS-SPORT | | Mishra[96] |
| | | | Density (g/mL) | 237 | | | |
| | | | Aromatic (wt%) | 237 | | | |
| | | | Viscosity (cst) | 237 | | | |
| 2021 | NIR | Diesel | Viscosity (cst) | 67 | PLS | mean centering | Hradecká[97] |
| | | | CFPP (°C) | 64 | | | |
| | | | Pour Point (°C) | 57 | | | |
| | | | Aromatics (wt%) | 70 | | | |
| | | | Sulfur content (ppm) | 50 | | 1st derivative | |
| 2022 | NIR | Diesel | Density (g/ml) | 243 | Automatic Model Construction | 1st derivative | Yu[98] |
| | | | | 53 | | mean centering | |
| 2022 | NIR | Diesel | Freezing point (°C) | 389 | IGWO | mean centering | Liu[99] |
| | | | Density (g/ml) | | | | |
| | | | Viscosity (cSt) | | | | |
| | | | Boiling Point (°C) | | | | |
| | | | Cetane Number | | | | |
| | | | Aromatics (wt%) | | | | |
| 2022 | NMR | Diesel | Diesel adulteration | 117 | Data fusion | normalization | Aguiar[100] |

From the information reported in Table 4, it could be inferred that middle distillates are the most researched cuts, occupying 68% of the studies (51% diesel and 17% kerosene). On the other hand, gasoline and lubricant oils share the same number of studies (≈15%). The remaining 2% of the research was focused on light atmospheric gas oil (LGO). Forty-nine fuel properties were studied among all the research collected, corresponding to 11 properties of gasoline, 16 of diesel, 15 of kerosene, and 7 of lubricant oil. Regarding middle distillates, the most studied properties for diesel were the cetane number and density, while the aromatics content and the distillation curve for kerosene.

58% of the 58 studies consolidated in Table 4 used NIR spectroscopy to develop the predictive models. MIR accounted for 30% of the studies, while the remaining 12% employed NMR spectra. For model development, the regression method most used was the partial least squares (PLS), representing 57% of utilization compared to other methods, including those related to machine learning, representing only 18% of implementation. Regarding spectroscopic information preprocessing, the two most frequently employed methods were the Savitzky-Golay (SavGol) derivative and the standard normal variate (SNV). A further fact extracted from the information analyzed was that the figure of merit most used for model evaluation was the root mean square error of cross-validation (RMSECV), using the leave-one-out method.

An interesting fact to extract from the compilation made by Moro et al.,[43] and the bibliographic analysis made in this thesis is that before 2009 all the studies used each spectroscopic technique independently to construct

the models. Since this year, the simultaneous use of information has begun to gain ground in developing chemometric models for property estimation in the oil and gas industry, mostly for crude oil analysis and to a lesser extent for fuels, especially middle distillates. The information discussed in the last paragraphs is summarized in Figure 3.

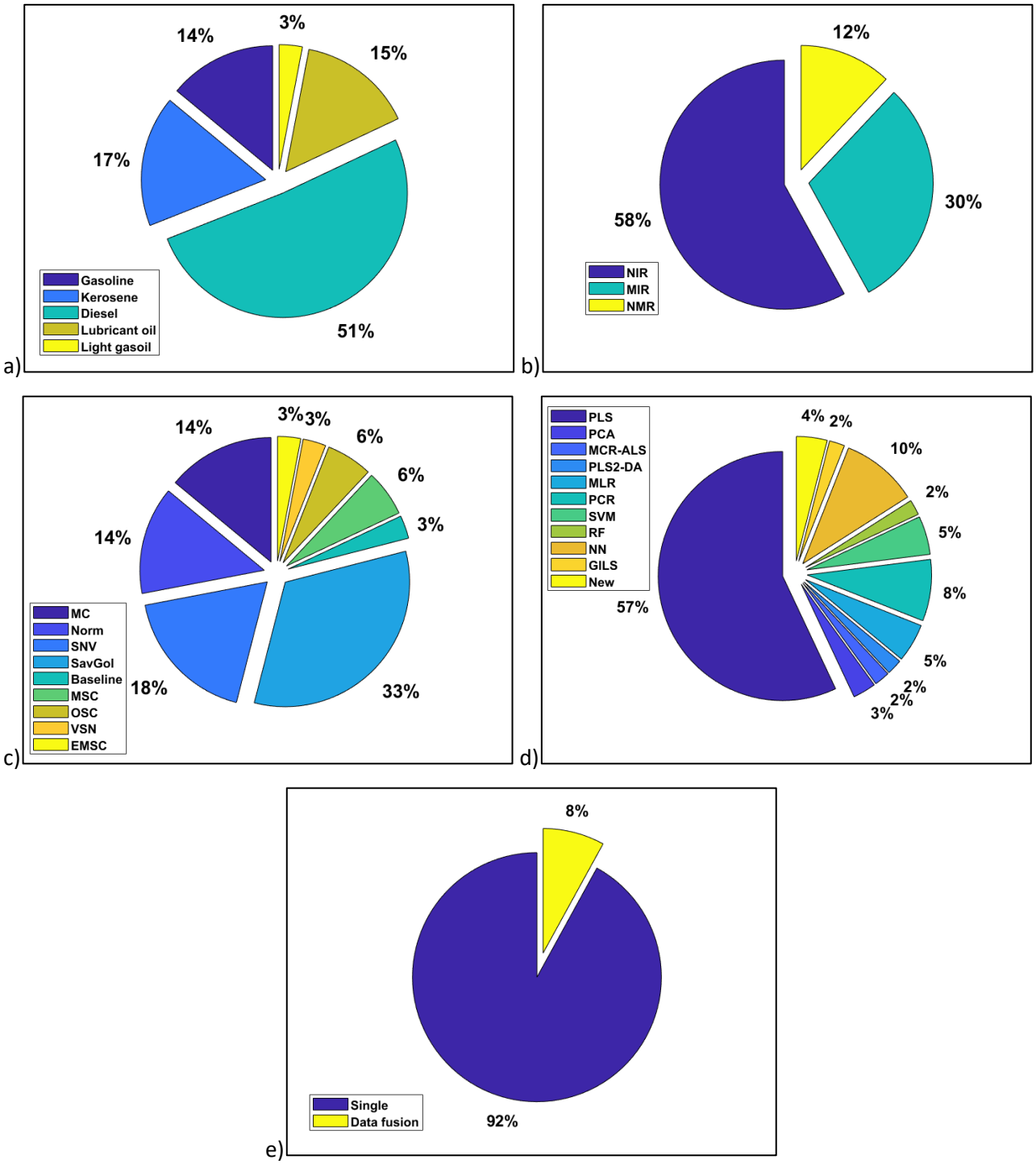



*Figure 3. Summary of the collection of chemometrics studies applied for fuel properties estimation. a) Product analyzed, b) Analytical technique employed, c) Preprocessing method employed, d) Regression method employed, e) Modelling approach employed*

Since the focus of this thesis is the estimation of middle distillate properties, a more detailed analysis of the studies involving these cuts was performed as follows.

### 2.2.1 Kerosene

Three properties of kerosene were studied during the development of this thesis, namely the cetane number, the flash point, and the pour point. The description of the kerosene cut, and the properties studied can be found in Chapter II, "Materials and methods".

The cetane number is generally measured using the ASTM D-613 standard[25]. This analysis is destructive, generally requires 500 ml of sample, and the reproducibility limits are defined as a function of the average measured value (see Table 2). For estimating this property, Zanier-Szydlowski et al.,[48] developed a 5-Latent Variable (LVs) PLS model from a database containing 90 NIR spectra acquired on diesel and kerosene samples. Their objective was to develop a comprehensive model to evaluate this property in the measurement range of middle distillate cuts. Using the mean center as the only preprocessing of the spectroscopic information, they were able to obtain a standard error of prediction (SEP) of 2.0, a value that is below the reproducibility limits of the reference method. An advantage evidenced in this study was the model application range (20 - 65) which is wider and covers more variability of the cetane number regarding the standard norm.

Through an internal research project at IFPEN, a PLS model for the kerosene cetane number estimation from NIR spectra acquired in kerosene samples was developed. The results were compared and validated against the ASTM D613 standard[25], showing prediction errors inferior to 1. This alternative for kerosene cetane number estimation is currently employed to characterize this stream. Going one step further in optimizing the response time to estimate the kerosene cetane number, it was demonstrated in an internal IFPEN study the potential of estimating this property from NMR spectra acquired on the total effluent obtained from the HCK process. From a database of 225 NMR spectra, a PLS model of 6 LVs with average prediction errors of 0.8 was obtained. The application range of this model is between 27.8 and 46.3, offering more flexibility in the estimation of low cetane number value compared to the standard method.

Concerning the flash point (FP), it is measured according to the ASTM D93-18 standard[26]. The reproducibility of this method is a function of the measured FP, and is defined by the formula ±0.071*FP. The FP of kerosene normally ranges between 35 and 65; therefore, the developed models should have prediction errors between 2.5 and 4.5. Using NIR spectra acquired on 29 kerosene samples, Garrigues et al.,[45] obtained a PLS model of 3 LVs for estimating the FP with a SEP of 2.1. Although the SEP is lower than the reproducibility stipulated by the standard norm, it must be considered that the samples used in the development of the model are only a few (29). In addition, this study does not report the range of application of the model, thus avoiding determining its versatility. Alternatively, Brouillette et al.,[82] used a more comprehensive database (92 samples). Using the first derivative SavGol with a third-order polynomial and an 11-point window as a preprocessing method, they obtained a PLS model of 5 LVs that can be applied in an estimation range

between 40 and 65. However, the RMSECV reported in this study exceeds the reproducibility limit of the reference method (>4) and has a low correlation squared coefficient ($r^2CV$) (0.53).

Smoke point (SP) has the least reported studies compared to the other two properties. This property is measured using ASTM D1322-12[27], and like the FP, the reproducibility is a function of the measured SP value given by the formula ±0.01651*(SP+30). Generally, kerosene has SP values between 15-45, which leads to expect regression models that present prediction errors around 1.2. In the same internal IFPEN study conducted on estimating the kerosene cetane number from the NMR acquired on total effluent samples, a PLS model of 8 LVs for predicting the kerosene SP was developed from 197 samples. The application range of this model is between 13 and 32, and its root mean squared error of prediction (RMSEP) is 1.1. Once again, the potential to reliably estimate kerosene properties from total effluent spectroscopic information is evident.

### 2.2.2 Diesel

In this thesis, the research work was focused on four properties of diesel, namely the cetane number and the cold flow properties (cloud point (CP), pour point (PP), and cold filter plugging point (CFPP)). The description of this cut and the properties studied are presented in detail in Chapter II, "Materials and methods".

The most researched diesel property is the cetane number. This property in diesel is measured using the same standard procedure employed in measuring the kerosene cetane number. As discussed previously, Zanier-Szydlowski et al.,[48] developed a model for cetane number prediction that covers diesel and kerosene cetane number ranges. In addition, in a homologous manner as for kerosene, at IFPEN it was developed a PLS model for predicting this property from NIR spectra acquired on diesel samples. This model validated against ASTM D613[25] is included in the model set used for estimating certain properties of middle distillates. With an application range between 40 and 60, Özdemir et al.,[59] developed a prediction model from the NIR spectra acquired on 245 diesel samples through a genetic regression algorithm associated with inverse least squares regression (GILS). The developed model shows an adequate performance by having a SEP of 2.1 and an $r^2$ of 0.86. Using the same analytical technique (NIR), Brouillette et al.[82] developed a 5-LVs PLS model for diesel cetane number prediction having a similar SEP (2.2) and range of application (43-57).

The most recent studies using NIR spectra in the diesel cetane number estimation are reported by Zhan et al. [101] and Barra et al. [94]. In the first study, a least squares-support vector machine (LS-SVM) regression model was developed, showing errors of calibration (1.8) and prediction (2.0) lower than the reproducibility of the reference method. However, the squared correlation coefficients of calibration ($r^2c$) and prediction ($r^2p$) were quite low (0.66). In the second study, using a PLS regression model with 8 LVs, diesel cetane number estimations were obtained with an RMESP around 0.5 and an $r^2p$ value higher than 0.9. Regarding this last study, it should be noted that the number of samples used for testing the model developed is low (10) with a narrow cetane number range (49-59).

Using a different analytical technique from the one discussed in previous paragraphs, Souza et al.,[76] present a study where estimation of the cetane number with errors below 0.7 is achieved using a PLS model of 4 LVs obtained from 60 NMR spectra acquired on diesel samples. While this is a rather promising result, it should be noted that the range of application of this model is limited (42-46.8). With similar results, Basu et al.,[53] propose an artificial neural network (ANN) model based on principal components using the NMR spectra acquired on 60 diesel samples. Unfortunately, the study does not report the application range of the model. Another study using the NMR acquired on diesel samples for predicting the cetane number of this product was developed by Santos et al.[78]. In this study, they estimate this property in diesel and blends of diesel with biodiesel, having errors lower than 0.8 and $r^2$ higher than 0.9. The application range of the PLS model developed is between 44 and 49.5.

Following the same internal study developed at IFPEN for kerosene cetane number estimation from the NMR spectra acquired on total effluent samples obtained from the HCK process reactors, an 8-LVs PLS model was developed for predicting this property in diesel. The developed model has the widest application range (30.5-70.7) regarding the studies discussed, including those developed with NIR spectra, and it provides estimates of this property with an RMSEP of 1.4.

In addition to the diesel cetane number estimation, some studies developed for predicting the diesel cold flow properties were analyzed. These properties have represented a challenge for developing chemometric models that allow their estimation reliably while having errors close to the reproducibility limits of the reference methods.

The CFPP is determined by NF EN 116[30], and its reproducibility limits are given by the formula ±3-0.06*CFPP. The lowest typical value of this property in diesel is -50°C, while the highest can be up to 15°C. Baptista et al.,[58] developed a PLS model of 3 LVs from NIR spectra acquired on 71 diesel samples. Using the second derivative of SG with a third-order polynomial and a 9-point window as a preprocessing method, they achieved an RMSEP of 1.1 for an estimation range between -14°C and 5°C. On the other hand, in a recent study reported by Hradecká et al.,[97] a PLS model of 10 LVs was obtained, allowing a wider range of application (from -47°C to 6 °C). As expected, the prediction error in this model (3.6) is higher than that reported by Baptista; however, this value is still close to the reproducibility limits of the reference method.

Another of the cold flow properties of diesel studies was the cloud point (CP). This property is measured using the ISO 3015 standard[28], which has a reproducibility of ±4. To estimate this property using NIR spectroscopy, we can highlight the works of Palou et al.,[89] and Brouillette et al.[82]. Both studies used the SavGol derivative as a preprocessing method. With a database of 278 diesel samples, Palou obtained a PLS model with 7 LVs that enables the CP estimation with an RMSEP of 1.15. This model has a range of application between -12.4°C and 2.2. In contrast, Brouillette obtained a PLS model with 5 LVs from NIR spectra acquired on 111 diesel samples. This model offers a wider range of CP estimation (from -25°C to 15°C); however, the RMSECV (3.6) is higher. Despite its higher RMSECV, its performance is acceptable compared to the standard

norm.

The third cold flow property of diesel is the pour point (PP). The standard norm used for measuring this property is the ASTM D5949[29], and it has a reproducibility of ±6. Once again, Brouillitte et al.,[82] succeeded in using a PLS model to obtain estimates of this property with errors close to the reproducibility limit of the reference method (5 vs. 6). The range of application of this model is between -28°C and -6°C. Hradecká et al.,[97] obtained a better-performing model with a wider application range (from -51°C to -7°C). Using 57 diesel samples, they developed a PLS model with 9 LVs that provides diesel PP estimation with an RMSECV of 3.6.

A very promising approach for estimating the cold flow properties of diesel, which was the foundation of the development of this thesis, is reflected in the internal IFPEN study where they succeed in predicting the three cold flow properties from NMR spectra acquired on total effluent samples from the HCK process. For the CFPP, a 10 LVs PLS model obtained from 254 NMR spectra estimates this property with an RMSEP of 2.61 and an application range between -32°C and -5°C. For the CP, 12 LVs of a PLS model are necessary to have estimates with an RMSEP around 2.3. This model has an application range between -40°C and -0.6°C. Finally, with a PLS model of 12 LVs, the PP can be estimated with an RMSEP of 2.8 and a model application range between -45°C and 3°C. As in the case of kerosene properties, this approach shows the potential for optimizing the response time in estimating middle distillate properties.

### 2.2.3 Real-time applications

Considering the need described in the previous section about achieving fast and reliable property estimations for process monitoring in real-time, this section also summarizes the main developments in implementing online regression models and some success cases applied to the oil industry.

The first step in process monitoring is to ensure that the chosen analytical method can produce results in real-time. As mentioned before, NIR spectroscopy meets this requirement, and its use in this type of application (online measurement) is not unknown. In the pharmaceutical industry, NIR spectrum online measurement for monitoring product quality is a growing practice. Its effectiveness and reliability have been demonstrated by several authors [102–104].

In the oil and gas industry, online measurement of the NIR spectrum has different applications, such as the measurement of physicochemical properties of crude oil (viscosity, density, metals, among others) for the preparation of blends[105]. Similarly, Kim et al. [50] used online NIR measurement to classify oil products in real-time with an error of less than 6% by combining PCA and Bayesian classifiers. The study conducted by Parisi et al.[44] is related to this subject since they sought to determine online fuel (gasoline and diesel) quality parameters such as Research Octane Number (RON), Motor Octane Number (MON), cetane number, and distribution of Paraffins Olefins Naphthenes and Aromatics (PONA), using NIR spectroscopy. Their results concluded that the NIR provides a reliable alternative for online determination of physicochemical

properties. It is important to highlight that all the studies described, recommend having appropriate sampling techniques and controlled conditions to ensure the reliable acquisition of the spectrum, avoiding adding noise to the models caused by external parameters. Other authors [106–110] used techniques other than NIR for predicting stream properties, including those in the domain of crude oil, showing interesting results in the use of the signal measured online. Each of these studies differs from the others in the regression method used. Due to the advantage that NIR has over NMR regarding its use in online (real-time) measurement applications, this method was the most exploited in this thesis.

As can be evidenced in the analysis presented, the use of spectroscopic analytical techniques as an alternative for estimating the middle distillate properties is of great interest and application. Nevertheless, it is worth highlighting that all the studies compiled in this literature analysis, except for the studies conducted at the IFPEN, estimate fuel properties from spectroscopic information acquired on them. Namely, diesel properties are estimated from NIR, or NMR spectra acquired on the diesel. At the time of this thesis development, no scientific publications that employed the approach of predicting middle distillate properties from spectral information of other related streams were found.

### 2.2.4 Conclusions

The hydrocracking process is of great importance for obtaining valuable products such as middle distillates (kerosene and diesel) from low-value streams such as the residues of crude oil distillation. Research in this process is vital to investigate the influence of different parameters such as feed quality, catalyst characteristics, and operating conditions (pressure, temperature, residence time) on product quality to determine the best process operating configurations. However, due to the extensive flexibility of the process, its research is a time- and cost-consuming task. Therefore, the challenges in optimizing this task to be more efficient (lower costs and shorter response time) are increasing.

Real-time estimation of middle distillates properties is an alternative for optimizing the response time. The adaptability and capabilities of the NIR spectroscopy make it suitable for achieving the analysis time reduction. In the oil & gas industry, the development of chemometric models from NIR spectra for estimating the physicochemical properties of these cuts has shown a promising outcoming. Even some process monitoring applications have been implemented, but none of them are in the hydrocracking process research field.

The accurate identification and understanding of the parameters affecting the process are crucial in developing prediction models for evaluating their variability. This task can be facilitated using additional information, such as NMR and process variables, that contributes with more detailed information about the sample's nature, improving the model performance. While some studies have already been developed using simultaneously analytical information from different sources, none involve middle distillates.

Developments made in estimating middle distillate properties from spectral information have always been based on the analytical information acquired on these cuts. Whereas these developments have shown a reduction in response time in the property estimation, their application remains dependent on the physical availability of the cut to be analyzed. Therefore, these developments still maintain the time constraint given by the atmospheric distillation to obtain the middle distillates. Moreover, most of the obtained models do not exhibit sufficient robustness to ensure their long-term effectiveness since only 10% of them correct the influence that external parameters may have on the quality of the spectral information and the model's performance. Only one study performed robust modelling as a strategy to correct the influence of external parameters.

An interesting approach, which solves the time constraint given by the distillation of the total effluent, is the one preliminarily investigated at IFPEN. This approach uses spectroscopic information acquired on the total effluent to estimate the properties of the middle distillates, thus reducing the need for the distillation procedure. Although the results are quite encouraging, the analytical technique employed in this first development (NMR) limits its application to real-time process monitoring. Predicting cut properties from the physical characteristics of a stream without going through distillation is a relatively new and growing approach that can make a significant contribution to process optimization using appropriate chemometric techniques.

Finally, there is no development in the oil industry related to the online prediction of the middle distillates properties from spectroscopic information of the total effluent obtained from the catalytic conversion process reactors.

# 3. Research problem

Considering the conclusions outlined, the main objective of this thesis is to develop reliable multivariate models that evaluate in real-time the different parameters affecting the hydrocracking process by predicting the physicochemical properties of middle distillates considering the synergy that may exist between the spectroscopic information of the total effluent and the HCK process variables. Furthermore, based on the advantages previously described related to the versatility and time of spectrum acquisition, NIR spectroscopy was chosen as the main method for developing and implementing chemometric models in the online characterization of middle distillates. To achieve the objective described, three main research questions were proposed.

1. Is it feasible to predict middle distillates properties from NIR spectra acquired on the total effluent obtained from hydrocracking process reactors?
2. Including additional and descriptive information to the NIR spectra improves the model performance?

3. Can external parameters' influence on the NIR spectra quality be compensated/corrected to ensure an online reliable properties estimation over time?

The results obtained during the thesis development made it possible to produce five scientific articles to address the research questions raised. These articles were the basis for constructing the manuscript chapters. Chapter 3 presents the results to respond to the first research question. This chapter is based on publications 1 and 2, which address the problem of developing chemometric models that enable the middle distillate properties estimation from the spectral information of the total HCK effluent, considering the best preprocessing and regression method. Chapter 4, based on publications 3 and 4, aims to address the research question related to the models' performance improvement when using complementary information to the NIR spectra when calibrating the models. Chapter 5 finally addresses the issue of model robustness. This chapter, linked to paper 5, shows the alternatives employed to estimate the middle distillate properties reliably under different evaluation conditions. The findings in these chapters were validated by applying the models in two case studies. The results of this validation are presented in chapter 6. Finally, the conclusions and perspectives are presented.

# CHAPTER II
# Material and Methods

# Chapter II. Material and methods

This chapter describes the materials and methods used in both the experimental and modelling work. The information presented in this chapter is replicated in detail in the articles produced, both those already published and those in the process of publication.

## 1. Analytical approach

Two groups of samples were involved in the development of the thesis, the total effluent from the HCK process reactors and the middle distillates (diesel and kerosene) obtained from the distillation of the total effluent. This section describes the analytical approach used in each of these groups.

### 1.1 Total effluent

As a reminder, the total effluent is obtained when heavy oil residues, mainly VGO, are processed in the reactors of the HCK process. Before the thesis development, the total effluent obtained from processing 32 feedstocks in the IFPEN pilot plants in Solaize, France, for over six years (between 2013 and 2018) were gathered. These samples were the core of the thesis experimentation since the spectra employed in the properties modelling were acquired on them. The following subsections broadly describe these samples' origins, the standard laboratory analyses used in their characterization, and the equipment employed in the experimental work implemented in the thesis.

#### 1.1.1 Origins and standard analysis

During the thesis development, 294 samples were analyzed from different tests performed in the HCK pilot plants located at the IFPEN facilities in Solaize, France. Table 5 shows the diversity of the experimental conditions used in these pilot plants for producing the total effluent samples analyzed.

*Table 5. Diversity of the operating parameters of IFPEN pilot plants in obtaining 294 total effluent samples*

| Process operating parameter | Number of parameter changes |
|---|---|
| Feedstock quality | 32 |
| Pressure | 8 |
| Temperature | 5 |
| LSHV | 10 |
| Catalytic system | 17 |

Table 6 summarizes the properties measured on the feedstocks used in the test conducted. In addition, this table shows four statistical parameters calculated on the information gathered to validate their physicochemical variability.

*Table 6. Summary of the variability of the properties measured on the 32 HCK feedstocks*

| Property | Standard norm | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|---|
| Density (g/ml) | ASTM D1218 - 12[21] | 0.8457 | 0.9837 | 0.9136 | 0.02662 |
| Refractive index | ASTM D1747[22] | 1.4442 | 1.5318 | 1.4853 | 0.02004 |
| Viscosity @ 70°C (cst) | ASTM D445-97[17] | 3 | 21 | 8 | 3.0 |
| Viscosity @ 100°C (cst) | | 6 | 76 | 21 | 11.2 |
| Sulphur (wt%) | ISO 20846[18] | 7E-04 | 3.5 | 1.1 | 1.15 |
| Nitrogen (ppm) | ASTM D5291[19] | 2 | 4825 | 1161 | 1143.8 |
| Hydrogen (wt%) | | 10.6 | 13.8 | 12.5 | 0.75 |
| Aromatic Carbon (wt%) | ASTM D3238-95[20] | 4.4 | 36.0 | 15.5 | 7.57 |
| Paraffinic Carbon (wt%) | | 43.1 | 71.6 | 57.1 | 5.33 |
| Naphthenic Carbon (wt%) | | 8.3 | 56.9 | 28.1 | 9.76 |
| SimDist IBP(°C) | ASTM D7213-15[23] | 68.4 | 358.4 | 228.2 | 91.11 |
| SimDist T5(°C) | | 121.6 | 403.8 | 324.4 | 56.59 |
| SimDist T10(°C) | | 159.6 | 414.7 | 357.0 | 43.55 |
| SimDist T20(°C) | | 216.0 | 435.6 | 390.6 | 33.64 |
| SimDist T30(°C) | | 268.6 | 449.8 | 412.0 | 28.42 |
| SimDist T40(°C) | | 323.2 | 466.4 | 429.4 | 25.05 |
| SimDist T50(°C) | | 368.6 | 479.8 | 445.0 | 21.86 |
| SimDist T60(°C) | | 389.5 | 498.2 | 462.0 | 20.44 |
| SimDist T70(°C) | | 409.4 | 514.7 | 480.0 | 19.24 |
| SimDist T80(°C) | | 433.4 | 537.2 | 501.0 | 19.22 |
| SimDist T90(°C) | | 466.4 | 563.7 | 529.2 | 18.77 |
| SimDist T95(°C) | | 493.3 | 606.4 | 552.1 | 17.01 |
| SimDist FBP(°C) | | 558.1 | 685.7 | 611.0 | 17.92 |

In turn, Table 7 summarizes the variability of the pilot plant operating conditions used during the conducted experimental tests.

*Table 7. Pilot plant operating conditions summary*

| Parameter | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|
| Pressure (bar) | 30 | 160 | 121 | 29.7 |
| Temperature R1 (°C) | 350 | 415 | 385 | 14.6 |
| Temperature R2 (°C) | 370 | 420 | 392 | 13.1 |
| LSVH ($h^{-1}$) | 0.4 | 4.0 | 1.6 | 0.92 |
| HDT catalyst | Parameters CHDT1, CHDT2, CHDT3, CHDT4, CHDT5 | | | |
| HCK catalyst | Parameters CHCK1, CHCK2, CHCK3, CHCK4, CHCK5 | | | |

As shown in Table 7, the information regarding the catalytic system used was coded to respect the confidentiality agreements related to this type of information. Finally, the diversity of the total effluent samples can be observed in the properties reported in Table 8 and their color and opacity, as shown in Figure 4. Since diesel and kerosene cuts are embedded in the total effluent, Table 8 also shows the variability of these cuts yields (cut weight percentage that can be recovered from the total effluent).

*Table 8. Variability of the physicochemical properties of the 294 total effluent samples. SimDis IBP, T5 - T95, description: Simulated distillation to determine the temperatures to start the sample evaporation and to recover from 5% to 95% of sample distillate*

| Property | Standard norm | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|---|
| Density (g/ml) | ASTM D1218 - 12[21] | 0.7891 | 0.9368 | 0.8640 | 0.03992 |
| Refractive index | | 1.2701 | 1.4975 | 1.4559 | 0.02874 |
| SimDist IBP(°C) | ASTM D2887-19ae2[24] | 60.2 | 280.4 | 119.5 | 46.82 |
| SimDist T5(°C) | | 69.1 | 376.2 | 207.1 | 89.55 |
| SimDist T10(°C) | | 90.0 | 401.0 | 242.7 | 95.86 |
| SimDist T20(°C) | | 117.8 | 426.3 | 285.3 | 95.84 |
| SimDist T30(°C) | | 143.9 | 442.0 | 316.2 | 92.49 |
| SimDist T40(°C) | | 168.8 | 456.9 | 343.0 | 87.17 |
| SimDist T50(°C) | | 194.3 | 472.5 | 368.8 | 80.78 |
| SimDist T60(°C) | | 219.9 | 489.1 | 394.7 | 73.07 |
| SimDist T70(°C) | | 250.6 | 506.6 | 422.1 | 64.91 |
| SimDist T80(°C) | | 282.8 | 528.8 | 451.6 | 55.92 |
| SimDist T90(°C) | | 329.0 | 554.0 | 488.5 | 46.72 |
| SimDist T95(°C) | | 367.2 | 585.3 | 516.4 | 41.35 |
| SimDist FBP(°C) | | 472.0 | 660.6 | 581.3 | 27.87 |
| Conversion in 370°C$^+$ (wt%) | - | 3.4 | 96.0 | 38.7 | 28.28 |
| Kerosene yield (wt%) | - | 0.4 | 48.8 | 15.9 | 14.41 |
| Diesel yield (wt%) | - | 4.9 | 55.7 | 29.6 | 14.35 |

*Figure 4. Variability of the color and opacity of total effluent samples obtained from the HCK reactors*

### 1.1.2 Near infra-red (NIR)

Two spectrometers were used to acquire NIR spectra on the total effluent samples. The first was a Fourier Transform Near-Infrared spectrometer (FT-NIR) MATRIX-F (Bruker, Optik GmbH, Ettligen - Germany), which with a resolution of 4 $cm^{-1}$ recorded 4148 wavenumbers within the range of 12000 - 4000 $cm^{-1}$. 32 scans were averaged in each acquisition to obtain the final spectrum. For acquiring absorbance spectra, the spectrometer system was equipped with an immersion transflectance Falcata Lab6 probe (Hellma GmbH & Co. KG, Müllheim – Germany) with an optical path fixed at 2 mm withstanding temperatures ranging from -40 °C to 200 °C. The software used with the spectrometer was OVP (OPUS Validation Program - Bruker, Optik GmbH, Ettligen - Germany) which automatically performs a series of analyses of the instrument's performance, evaluates them and ensures that it is operating within specifications (See Figure 5).


*Figure 5. Experimental apparatus employed in the NIR spectra acquisition (Bruker spectrometer)*

The second spectrometer was a NIRS XDS Process Analyzer (Metrohm, Villebon - France). This spectrometer, having a resolution of 0.5 nm, recorded wavelengths within the 800 - 2200 nm spectral range. The final spectrum obtained in each acquisition was the average of 32 scans performed on the samples. The software used with the spectrometer was VISION (Metrohm, Villebon - France). During the thesis, this spectrometer was employed to acquire spectra at steady-state and dynamic conditions. A Falcata Lab6 immersion probe (Hellma GmbH & Co. KG, Müllheim - Germany) with two optical lengths of 1 and 2 mm was used for the steady-state acquisition (Figure 6 green square). For the dynamic acquisition, a transmission Flow cell of 1/4" OD tube and optical length of 1 mm was used (Figure 7).

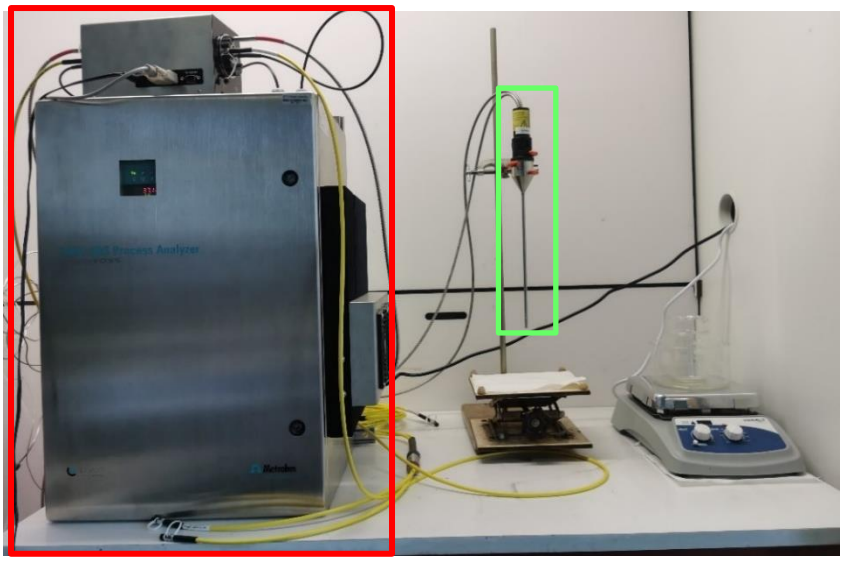


*Figure 6. Experimental apparatus employed in the NIR spectra acquisition. Red square → NIR XDS spectrometer, green square → immersion probe*


*Figure 7. Transmittance Flow cell*

To preserve the total effluent samples' integrity, they were stored in a cold room (temperature below 0°C) and remained there until the experimental test. Once the spectral acquisition was completed, the samples were returned to the cold storing room. For this reason, prior to any acquisition of NIR spectra, the total effluent samples were heated in closed flasks at 60°C for one hour in a water bath and shaken manually to ensure their liquid state and homogeneity.

### 1.1.3 Nuclear Magnetic Resonance (NMR)

Although the development of the thesis focused on the exploitation of NIR spectroscopy, the use of NMR was crucial in solving one of the research questions. The NMR spectra of the total effluent samples were already available at the beginning of the research; nevertheless, the experimental setup employed in acquiring these spectra is described below.

For the NMR spectra acquisition, 250µl of total effluent were mixed with 250µl of $CDCl_3$ and 0,3mg of $Fe(acac)_3$. $^{13}C$ NMR spectra were recorded at 50°C on a Bruker Advance 600 MHz spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at 150.9 MHz using a 5 mm QNP probe (time-domain 128k, 60° pulse, proton decoupling, acquisition time 56 min, relaxation delay 5 s, 512 scans). Zero filling and exponential line broadening (1 Hz) were applied before the Fourier transform. The spectra were accurately phased and baseline adjusted. The $^{13}C$ NMR chemical shift of chloroform-d was set to 76.9 ppm as an internal standard. Once again, to ensure the homogeneity of the samples, they were heated at 70°C and manually shaken before performing the NMR analysis.

## 1.2 Middle distillates

For this thesis, the experimental work was limited to the spectral acquisition of the total effluent. The analytical information of the middle distillates used in developing the chemometric models was already available at the beginning of the thesis. Nevertheless, it is worthwhile to describe the materials and methods used to obtain and characterize these samples. Therefore, the analytical approach used in recovering the samples and obtaining the studied properties is described below.

### 1.2.1 Samples recovery

Based on the ASTM D2892-20 standard[111], the kerosene and diesel samples were recovered from the distillation of the total effluent samples. For the kerosene cut, it was used an initial boiling point (IBP) between 150 °C and 180 °C and a final boiling point (FBP) between 225 °C and 250 °C. For the diesel cut, it was used an IBP temperature between 250 and 275°C and a FBP temperature between 340 and 370°C.

### 1.2.2 Kerosene

The kerosene is the lightest cut of the middle distillates. The most relevant properties in the characterization of this product are the cetane number[25], the flash point[26], and the smoke point[27].

The cetane number determines the ignitability of the sample using a standardized engine and reference fuel. The cetane number is determined by comparing the ignition time of a mixture of cetane and hepta-methyl-nonane, having the same ignition time delay as the tested sample. The kerosene cetane number was measured on each diesel sample recovered using an IFPEN internal method, which estimates this property from diesel NIR spectra through a PLS model based on Zanier-Szydlowski et al. work [48], with a larger database and equivalent performance. The internal method outlined was developed using the cetane numbers measured using the ASTMD613-01 standard[25] analysis as the reference method and validated against the reproducibility limits defined by this norm. Table 9 summarizes the available analytical information for this property. This table shows the variability of the cetane number using four statistical parameters. Also reported in this table is the analytical method to measure the property, its respective reproducibility value, and the amount of data available concerning the total effluent samples collected.

The flash point (FP) is the lowest temperature at which sufficient vapor is emitted to form a flammable mixture in the air at standard atmospheric pressure. This property was measured by heating the sample at specific temperatures and under controlled conditions. At each temperature tested, a spark is applied until a flame is generated. The measurement of this property was performed using the ASTM D93-18 standard[26]. Table 9 also summarizes the available analytical information for this property.

Finally, the smoke point (SP) in the kerosene samples was measured using the standard ASTM D1322-12[27]. This property measures the tendency of a fuel to generate smoke when burned. The smoke point is established by "the maximum height, in millimeters, of a smokeless flame of fuel burned in a wick-fed lamp of specified design"[27]. The higher the smoke point, the better the quality of fuel. Table 9 shows the consolidated information on this property.

*Table 9. Summary of the variability of the properties measured on kerosene samples and their respective reference method*

| Property | Minimum | Maximum | Mean | Standard deviation | Number of data available | Reference Method | Reproducibility |
|---|---|---|---|---|---|---|---|
| Cetane number | 21.5 | 46.0 | 38.5 | 5.12 | 93 | IFPEN method | ±3.6 |
| Flash Point (FP) (°C) | 42.0 | 56.5 | 52.9 | 2.53 | 35 | ASTM D93-18 | ±0.071*FP |
| Smoke Point (SP) (mm) | 13.3 | 34.0 | 24.5 | 3.91 | 82 | ISO 3014 | ±3 |

### 1.2.3 Diesel

The properties analyzed in the diesel cut were the cetane number described previously and the cold flow properties, namely, the pour point (PP)[29], the cloud point (CP)[28], and the cold filter plugging point (CFPP)[30]. The CP is the most considered parameter for the formulation of diesel fuel[112]. This property specifies the temperature when the first paraffin or wax crystals appear, causing the fuel to turn cloudy. This measurement is done by cooling the diesel sample according to a given cooling curve and checking it periodically until the

first wax crystal is deposited at the bottom of the vessel. The standard norm used to measure this property is the ISO 3015[28]. In turn, the PP is measured using the ASTM D5949[29]. This method uses an optical device to measure the fluidity of the sample by applying a burst of nitrogen gas while the sample is being cooled. The PP seeks to determine the temperature at which the diesel stops flowing. Finally, the CFPP establishes the temperature at which the crystallized wax begins to plug a standardized filter arrangement (simulating the fuel filter in a diesel engine) in such a way as to hinder the fuel flow. The NF EN 116[30] is the standard used to measure this property.

Table 10 summarizes the statistical parameters calculated based on the information available for the diesel properties.

*Table 10. Summary of the variability of the properties measured on diesel samples, and their respective reference method*

| Property | Minimum | Maximum | Mean | Standard deviation | Number of data available | Reference Method | Reproducibility |
|---|---|---|---|---|---|---|---|
| Cetane number | 30.3 | 69.5 | 52.0 | 10.72 | 131 | IFPEN method | ±3.6 |
| Cloud Point (CP) (°C) | -31.0 | 13.0 | -15.6 | 7.93 | 139 | ISO 3015 | ±4 |
| Pour Point (PP) (°C) | -42.0 | 15.0 | -18.8 | 11.34 | 104 | ASTM D5949 | ±6 |
| Cold Filter Plugging Point (CFPP) (°C) | -29.0 | 5.0 | -13.1 | 8.71 | 127 | NF EN 116 | ±3-0.06*CFPP |

Table 9 and Table 10 evidence that the number of analytical information available for the middle distillates is not the same for all properties as it varies according to the previous studies' requirements and experimental planning. However, the analytical variability is sufficiently informative to be representative of the different scenarios that may arise in the HCK process research.

# 2. Modelling approach

The thesis work focused on exploiting the available and acquired spectroscopic information to develop regression models enabling the estimation of middle distillate properties from the analytical information of the total effluent. The different chemometric methods used in pursuing this objective are presented in this section. The detailed description regarding the use of the methods stated in this section can be found in the following chapters.

## 2.1 Data analysis methods

The analysis of the data before model development is an important task that contributes to the understanding and definition of the suitability of the information and the different parameters that may affect the data quality. Although different chemometric methods can be used for a preliminary data analysis,

the research limited these methods to the principal component analysis (PCA)[113], the hierarchical cluster analysis (HCA)[114], and the Q residual and Hotelling $T^2$ tests [115].

The use of the PCA analysis had a twofold objective. First, this analysis was used to evaluate the variance of the total effluent samples, the distribution of the calibration and test data sets, and the relationship that might exist between the spectroscopic information and the different parameters that could affect both the quality of the spectra (type of experimental apparatus and the spectral acquisition conditions used) and the properties of the middle distillates studied (process variables). Second, This analysis was used in conjunction with the HCA analysis to evaluate the effectiveness of some preprocessing methods in correcting for the impact of consecutive and repetitive spectrum acquisition (noise associated with lack of repeatability and reproducibility) and sample temperature[116].

The combined analysis of the Q residual and Hotelling $T^2$ tests was used primarily to evaluate the consistency of each total effluent sample regarding the properties evaluated. In addition, this analysis allowed identifying potential anomalous data in the datasets (calibration or test), helping to identify the cause of such behavior.

## 2.2 Preprocessing methods

Preprocessing the spectral data is one of the first steps in constructing a chemometric model. Hence, an analysis was conducted to determine the best preprocessing scheme for each middle distillate property. This thesis analyzed 9 of the most common preprocessing methods applied to NIR spectra, summarized in Table 22, using an in-house MATLAB script. Each method and its possible combinations were evaluated based on the performance of different PLS regression models built using the root mean square error of cross-validation (RMSECV) as the figure of merit.

As mentioned in section 1.1.3 of this chapter, the NMR spectra of the total effluent samples collected were used in this thesis as complementary information to answer one of the research questions. However, unlike the NIR spectra, when NMR spectra were employed, the preprocessing applied was the same for all the properties studied. First, the NMR spectra were aligned using the Interval Correlation Optimized (icoshift) algorithm[117]. Subsequently, these spectra were preprocessed using the Savitzky-Golay smoothing (15-point window, polynomial order = 0)[118] and normalized (each variable is divided by the sum of the absolute values of all the variables for a given spectrum)[119].

*Table 11. Pre-processing method evaluated on the NIR spectra of the HCK total effluent*

| # | Category | Method | Acronym | Parameters |
|---|---|---|---|---|
| 1 | Normalization | Variable Sorting for Normalization[120] | VSN | Automatic calculation |
| 2 | | Standard Normal Variate[121] | SNV | |
| 3 | | Multiplicative Signal Correction[122] | MSC | Reference data = mean of data, whole spectral range |
| 4 | | Probabilistic Quotient Normalization[123] | PQN | |
| 5 | Filtering | Automatic Weighted Least Squares Baseline[124] | AWLS-B | |
| 6 | | Detrend[121] | Dt | Polynomial order (1-3) |
| 7 | | Extended Multiplicative Scatter/Signal Correction[125] | EMSC | Reference spectrum (basis to remove the scatter) = mean of each matrix generated, polynomial order = (1-4), whole spectral range, algorithm (CLS, ILS)* |
| 8 | | Norris-Williams Derivation[126] | NW-D | Window points (9-25), gap size = (3-9), First order derivation |
| 9 | | Savitsky-Golay Derivative[118] | SG-D | Window points (9-25), polynomial order = (1-4), derivative order (1-4) |

** CLS = Classical Least Squares, ILS = Inverse Least Squares.*

## 2.3 Regression methods

From the analytical information described in section 1 of this chapter, a master database containing two data types, namely, low multivariate data (operational information) and high multivariate data (spectroscopic techniques), was consolidated. Both types of data were employed in the applied modelling approach, divided into two main groups: single data and data fusion modelling. The different regression methods applied in each modelling group are presented below.

### 2.3.1 Single data modelling

This type of modelling was used to calibrate NIR and NMR models from the spectra acquired on the total effluents, using each of these data blocks independently. Given the advances shown in recent years in the computational field for processing and analyzing multivariate data, different regression methods have been proposed and are currently the subject of research and development. For this reason, the performance of 3 different multivariate regression methods was evaluated regarding PLS performance, particularly those that offer a solution to the possible nonlinearities that the collected information may present. Table 12 shows the methods evaluated.

*Table 12. Regression methods employed in the single modelling approach*

| Method | Acronym |
|---|---|
| Partial Least Squares | PLS[127] |
| Support Vector Machine | SVM[128] |
| Artificial Neural Network | ANN[129] |
| Locally Weighted Regression | LWR[130] |

### 2.3.2 Data fusion modelling

The continuous increase in the generation of information from different sources that can describe the sample physicochemical behavior has led to the simultaneous use of data in developing models. This type of data manipulation employs different strategies known as fusion levels (low-, mid-, and high-level) [131, 132]. The low-level fusion consists of using the information from the blocks directly in the development of the model either by simple concatenation of the blocks or using decomposition or factorization methods on one block regarding another [133]. At the mid-level fusion, a feature extraction step from each dataset is performed first through statistical analyses such as PCA and PLS for their later fusion by simple concatenation [134]. Finally, the high-level fusion combines the decisions or results obtained from developed prediction models separately with each data block [135]. Table 13 summarizes the methods used to evaluate the different levels of data fusion described.

*Table 13. Regression methods employed in the data fusion modelling approach*

| Method | Acronym | Fusion-Level used |
|---|---|---|
| Sequential Orthogonalised PLS | SO-PLS[136] | Low |
| Response-Oriented Sequential Alternation | ROSA[137] | |
| Principal Component Analysis | PCA | Mid |
| Partial Least Suares | PLS | |
| Multiple Linear Regression | MLR | High |

## 2.4 Variable selection methods

When generating a model using either univariate or multivariate data, some variables or characteristics might not have relevant information about the property being studied. For this reason, it is sometimes desirable to select the most descriptive variables and reduce their number as input to the model, either to optimize machine consumption, reduce costs in the use of sensors capturing the data, or improve model performance. This thesis conducted a variable selection analysis to determine the descriptors that would most impact estimating the properties studied. The purpose was twofold: to improve the predictive performance of the models and to understand the impact of the process variables and macroscopic characterization of the feedstock on the quality of the middle distillates. This section outlines the variable selection methods used in both univariate and multivariate data.

### 2.4.1 Low multivariate data

The methods shown in Table 14 were used to perform the variable selection on low multivariate data blocks. Although the VIP, SR, CovSel and SO-Covsel methods are mostly applied to multivariate data, they were applied to univariate data to evaluate their performance in fusing spectral and process data.

*Table 14. Variable selection methods applied to low multivariate data block*

| Method | Acronym | Parameters |
|---|---|---|
| Variable Importance in Projection | VIP[138] | Automatic feature selection |
| Selectivity Ratio | SR[139] | |
| Least Absolute Shrinkage and Selection Operator | LASSO[140] | Alpha tuning [0.01,0.07,0.05, 0.1, 1,2, 3, 5, 10] |
| Genetic Algorithm | GA[141] | Window width = 3, mutation rate = 0.005, 30% initial terms, Convergence = 50%<br>Algorithm = MLR, 15 runs |
| Recursive Feature Elimination | RFE[142] | Features tuning [9-22] |
| eXtreme Gradient Boosting Feature Selection | XGBoost_FS | Algorithm = gblinear, automatic feature selection |
| Sequential Forward Selection | SFS[143] | Algorithm = MLR, automatic feature selection |
| Sequential Backward Selection | SBS | |
| Sequential Forward Floating Selection | SFFS[144] | |
| Sequiential Backward Floating Selection | SBFS | |
| Covariance Selection | CovSel[145] | Features tuning [9-22] |
| Sequential Orthogonalised CovSel | SO-Covsel[146] | Features tuning: Automatic and [9-22] |

### 2.4.2 Multivariate data

The main goal of performing the variable selection analysis on multivariate data (NIR spectra) was to improve the performance of the developed models. The methods applied on this data type are summarized in Table 15.

*Table 15. Variable selection methods applied to multivariate data*

| Method | Acronym | Parameters |
|---|---|---|
| Variable Importance in Projection | VIP | Automatic |
| Selectivity Ratio | SR | |
| Genetic Algorithm | GA | Window width = 50, mutation rate = 0.005, 30% initial terms, Convergence = 50% Algorithm = PLS (20LVs), 15 runs |
| Forward interval PLS | F-iPLS[147] | Interval size [25,50,100,200] |
| Backward interval PLS | B-iPLS | |
| recursive PLS | rPLS[148] | Max. iteration = 500, Max. LVs = 20 |
| Covariance Selection | CovSel | Features tuning [25,50,100,200,400,800] |
| Sequential Orthogonalised CovSel | SO-Covsel | Features tuning: Automatic |

## 2.5 Methods for external parameters influence correction

The validity of a model relies on its performance over time; however, different parameters can impact the quality of the spectrum used as input to the model, affecting its reliability and rendering it obsolete and useless. There are four strategies to correct or compensate these parameters' incidence on the spectra: a priori correction, model correction, a posteriori correction, and robust modelling[149]. Although different methods can be used in each strategy, in this thesis, the evaluation was limited to the methods summarized in Table 16.

*Table 16. Methods for external parameters influence correction*

| Method | Acronym | Parameters | Associated strategy |
|---|---|---|---|
| Piecewise Direct Standardization | PDS[150] | Window [3 - 15] | *a priori* |
| Partial Least Squares | PLS | | Model Correction |
| Bias and Slope Correction | BSC[151] | | *a posteriori* |
| External Parameter Orthogonalisation | EPO[152] | EPO Components = [1-10] | Robust modelling |
| Dynamic Orthogonal Projection | DOP[153] | | |

## 2.6 Model evaluation criteria

As has been emphasized throughout this manuscript, the major motivation for this thesis, and hence its scope, is the development of robust models for predicting middle distillates properties from the spectroscopic data of the total effluent of the HCK process. As seen in the previous sections of this chapter, the volume of information available and the methods employed to develop the models are considerable. Consequently, it is necessary to establish evaluation criteria to select models with the best performance and maximum robustness.

Throughout the development of the thesis, the Root Mean Square Error of Calibration (RMESC), Cross-Validation (RMSECV), and Prediction (RMSEP) were systematically calculated on each model generated. These statistical parameters were the main criteria for evaluating the models and are summarized in Table 17. Broadly speaking, the RMSEC measures how well the calibration data fit the model generated using all the points from the same calibration set. This value gives a general idea of how well the model performs with samples having identical characteristics to those used in the development of the model, leading to erroneous conclusions if only this statistical parameter were used in the evaluation. Therefore, RMSECV was calculated to have a more representative analysis of the model's predictive potential with future samples. This error is estimated from an internal validation using different subsets of data defined from the calibration set according to the selected validation method, being the Venetian blind 10-fold the one used in this thesis.

The RMSEC and RMSECV played a key role in selecting one model over another. The model with the lowest RMSECV was selected in the comparative evaluation as long as the RMSECV/RMSEC ratio did not exceed 1.7. This parameter, defined empirically in previous studies, was intended to avoid overtraining the models. This criterion was also applied in the definition of the LVs to be retained in the PLS models. Finally, the best-performing models were tested using the independent validation set (different from the calibration set). The RMSEP, bias, and standard error of prediction (SEP) were calculated to evaluate the model performance on new samples.

Another complementary statistical parameter used in the evaluation of the regression models was the squared correlation coefficient ($r^2$) calculated between the measured value and the predicted value in both the cross-validation ($r^2CV$) and the prediction of the model ($r^2P$). Summarized in Table 17, this parameter helps to evaluate how well the variation of the studied properties is explained through the model.

Lastly, the percentage of predicted samples with residual values less than or equal to the reproducibility of the standard reference norms was calculated. This parameter was named "percentage of effectiveness" and was used to validate the models as an alternative in estimating middle distillate properties.

*Table 17. Statistical parameters for model performance evaluation*

| Parameter | Equation |
|---|---|
| Standard Error | $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i - Bias)^2}{n-1}}$ |
| Bias | $Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$ |
| Root Mean Square Error | $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$ |
| Squared correlation coefficient | $r^2 = r(\hat{y}, y)^2$ |
| Percentage of effectiveness | $\%eff = \frac{\sum_{i=1}^{n} 1_{(\hat{y}_i - y_i) \leq rep}}{n}$ |

$n$: *number of samples evaluated*
$\hat{y}_i$: *predicted value of sample i*
$y_i$: *measured value of sample i*
$rep$: *reproducibility limit of the reference method.*

The software employed in the development of the thesis were PLS_Toolbox V.8.9 (Eigenvector Research Inc. Wenatchee, WA, USA), MATLAB V.2020b (The MathWorks, Inc., Natick, MA, USA), and Python V3.6.

# CHAPTER III
# NIR Modelling

# Chapter III. NIR Modelling

The results discussed in this chapter are linked to the work reported in the scientific papers 1[116] and 2[154] (see appendices 1 and 2, respectively). The findings discussed in the papers, and this chapter, were the basis to address the first research question: "Is it feasible to predict middle distillates properties from NIR spectra acquired on the total effluent obtained from hydrocracking process reactors?" Accordingly, the work focused on seven properties of the middle distillates; four of the diesel cut (cetane number and cold flow properties (PP, CP, CFPP)) and three of the kerosene cut (cetane number, SP, FP).

The information presented in this chapter was divided into three sections. The first section shows the different modelling approaches employed, including the preprocessing of the NIR spectra[116] and the calibration of the predictive models[154]. It should be noted that this first section shows only in detail the work done on the diesel cetane number. Thus, the second section of this chapter summarizes the results for the other middle distillates properties. Finally, the third section presents the conclusions responding to the research question.

## 1. Diesel cetane number modelling

The first step was to generate the database employed in the diesel cetane number modelling. Chapter 2, "Materials and methods," details the protocol used to acquire the NIR spectra on the total effluent samples and measure the cetane number on the diesel samples. In summary, 98 total effluent samples were used for NIR spectra acquisition at constant conditions of sample temperature (60°C) and instrument optical length (2mm) to address the first research question. For modelling the diesel cetane number, a matrix x (Mx) containing the NIR spectra of the 98 samples and a matrix y (My) containing the corresponding measurements of diesel cetane number were generated.

The second step was to define the spectral range used. This analysis was based on the studies of Yalvac et al. [155] and Kelly et al. [156], resulting in choosing to work in the spectral region between 1110-2200 nm[154]. Next, the best preprocessing scheme applied to the NIR spectra was defined by evaluating the different methods summarized in chapter 2, Table 11. For the diesel cetane number, the best preprocessing scheme was the combination of the Standard Normal Variate (SNV) and the second derivative of Savitzky-Golay with a third polynomial order and a 23 window-point (SavGol[23,3,2]). Finally, the Mx and My matrices were divided into the calibration and test sets using the Kennard-Stone algorithm applied on the Mx. The second paper[154], found in appendix 2, gives a detailed description of this second step.

The final step was the NIR model calibration for predicting the diesel cetane number. Four regression methods (PLS, SVM, ANN, LWR) were evaluated using the datasets defined in the second step. Table 18

summarizes the calibration parameters defined in each method evaluated and the calculated statistical parameters, such as the errors (RMSEC, RMSECV, RMSEP) and the squared correlation coefficients ($r^2C$, $r^2CV$, $r^2P$), for assessing their performance. Section 2.6 of chapter 2 described that the RMSECV was calculated from an internal validation using different subsets of data defined from the calibration set using the Venetian blind 10-fold methodology. Figure 8 shows graphically the comparison between the different regression methods evaluated.

*Table 18. NIR models for diesel cetane number estimation comparison*

| Method | Configurations evaluated | Final configuration | RMSEC | RMSECV | RMSEP | Bias Pred | SEP | $r^2c$ | $r^2CV$ | $r^2P$ | RMSE (CV/C) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PLS | LVs (1-20) | 9 LVs | 1.3 | 2.2 | 2.0 | -0.6 | 1.9 | 0.986 | 0.959 | 0.955 | 1.7 |
| SVM<br>Gamma (10-6 - 10)<br>Cost (10-3 - 100)<br>E (1.0, 0.1, 0.01)<br>nu (0.2, 0.5, 0.8) | Kernel type (RBF) | | | | | | | | | | |
| | Compression (PCA 1-20 PCs) | 12 PCs | 1.2 | 2.4 | 1.9 | -0.9 | 1.7 | 0.988 | 0.950 | 0.962 | 2.0 |
| | Compression (PLS 1-20 LVs) | 7 LVs | 1.3 | 2.4 | 1.9 | -0.7 | 1.8 | 0.985 | 0.951 | 0.960 | 1.8 |
| | Kernel type (Linear) | | | | | | | | | | |
| | Compression (PCA 1-20 PCs) | 15 PCs | 1.4 | 2.5 | 2.2 | -1.0 | 2.0 | 0.970 | 0.943 | 0.951 | 1.8 |
| | Compression (PLS 1-20 LVs) | 10 LVs | 1.1 | 2.3 | 2.0 | -0.8 | 1.8 | 0.990 | 0.955 | 0.953 | 2.1 |
| ANN<br>Algorithm (BPN)<br>Learn rate (0.125)<br>Learn cycles (20) | Layers (2) Nodes (2-6) | | | | | | | | | | |
| | Compression (PCA 1-20 PCs) | 13 PCs<br>4 & 1 nodes | 0.9 | 2.8 | 1.8 | -0.5 | 1.7 | 0.993 | 0.896 | 0.962 | 3.1 |
| | Compression (PLS 1-20 LVs) | 12 LVs<br>4 & 2 nodes | 0.4 | 2.5 | 2.1 | -0.4 | 2.1 | 0.999 | 0.949 | 0.945 | 6.3 |
| LWR | Local points (10-30) | | | | | | | | | | |
| | PCR Algorithm (1-19 PCs) | 6 Pcs<br>24 points | 0.7 | 2.2 | 2.2 | -0.5 | 2.1 | 0.995 | 0.959 | 0.940 | 3.1 |
| | PLS Algorithm (1-19 LVs) | 5 LVs<br>30 points | 0.4 | 2.0 | 1.8 | -0.5 | 1.7 | 0.998 | 0.964 | 0.960 | 5.0 |

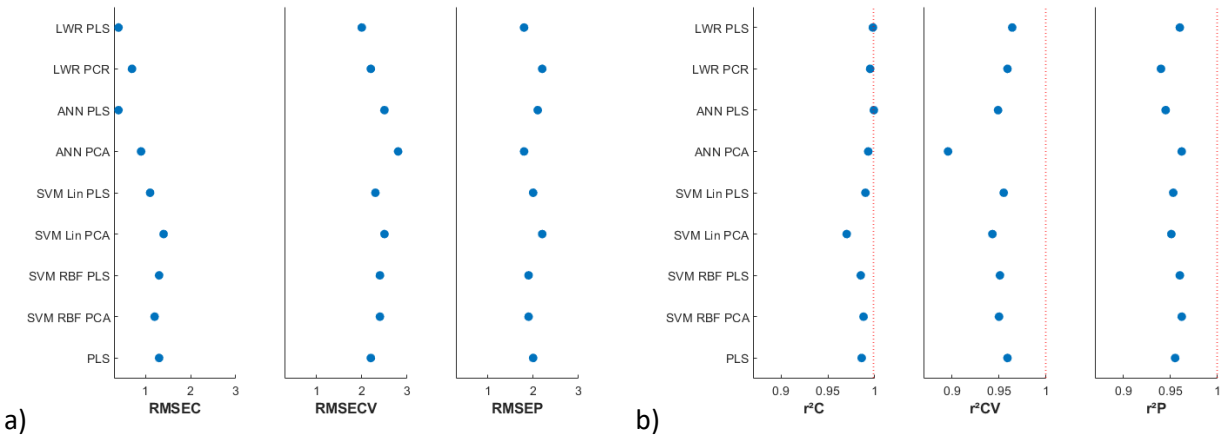



*Figure 8. Performance comparison of regression methods employed for diesel cetane number modelling a) model errors. b) model squared correlation coefficients*

From Figure 8, it is observed that the performance of the different models developed are rather similar when

comparing them using the RMSEP as a figure of merit. This closeness of performance is also reflected in the models' $r^2P$. Regarding the prediction bias, the PLS model presents a lower value (between 0.1 and 0.4 lower) than the different SVM models generated, ensuring a more accurate prediction. Compared to the ANN and LWR models, the prediction bias of the PLS model is slightly higher (between 0.1 and 0.2 higher). However, the impact of this value on the standard prediction error (SEP) is comparable for these three regression methods.

Alternatively, the RMSECV/RMSEC ratio was used as an evaluation criterion. This criterion, established empirically in previous modelling works, gives an idea of the obtained model consistency. The closer the ratio is to unity, the higher the consistency. Therefore, this criterion facilitates the identification of possible model overtraining. From Table 18, it is observed that the PLS and SVM models have the closest values of this criterion to unity. On the contrary, the ANN and LWR regression models present the highest values (>3), indicating lower consistency and a possible overtraining. The results observed when using these two regression methods can be attributed mainly to the limited database size used in the model calibration (67 samples).

Considering the previously discussed results, it is possible to establish that the PLS and SVM regression methods are the most adequate to develop models for estimating the diesel cetane number reliably. Nevertheless, the PLS method has an advantage over the SVM related to the model interpretability. While a black-box model is retrieved with the SVM method, the PLS method enables analyzing and establishing the coherence between the chemical information contained in the NIR spectrum of the total effluent and the diesel cetane number (see appendix 2[154]). Therefore, the model developed with the PLS method was defined as the one with the best performance.

In summary, a NIR PLS model with 9 LVs for the diesel cetane number estimation was developed from 67 spectra acquired on hydrocracked total effluent samples having a corresponding diesel cetane number range between 30.3 and 69.5. The reliability of the retained model was validated by comparing its performance against the reproducibility of the reference method employed to measure the diesel cetane number. For this property, the reference used was the internal IFPEN method, which has a reproducibility of ±3.6 (see chapter 2, section 1.5). The model validation was performed using the external test set composed of 31 total effluent spectra, with an associated diesel cetane number range from 37.3 to 69.3.

The primary criteria used to validate the reliability of the model were the RMSECV and RMSEP, whose values (2.2 & 2.0) were lower than the reference method reproducibility. A secondary criterion employed was the percentage of effectiveness in predicting samples from the test data set within these reproducibility limits. Figure 9a shows the parity plot between the measured and predicted value, showing that only one of the 31 samples predicted was outside the limits. Thus, this prediction of the diesel cetane number corresponds to 97% of effectiveness. Finally, two additional criteria used, which are not explicitly related to the reproducibility of the reference method, were the $r^2P$ and the distribution of the residual predicted values.

From Table 18, it can be inferred that the correlation between the measured and predicted value is satisfactory enough since the $r^2P$ is higher than 0.95. Regarding the prediction residuals, Figure 9b shows a semi-homogeneous distribution of these values over the entire evaluation range, validating the homoscedasticity of the model.

The analysis presented in the previous paragraph leads to the preliminary conclusion that reliable and reproducible estimation of the diesel cetane number using the NIR spectra of the HCK total effluent is feasible.

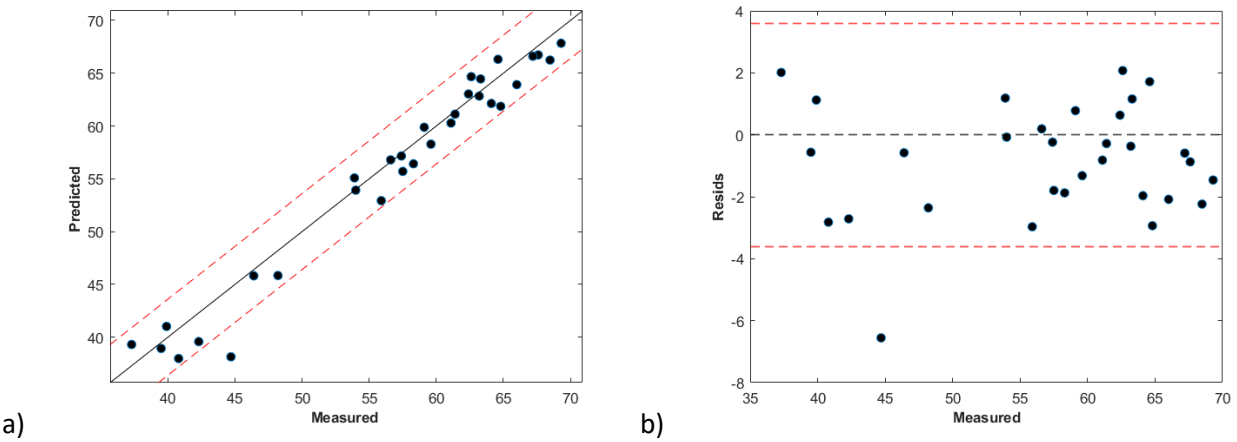



*Figure 9. a) Parity plot, b) residuals plot of PLS model for predicting the diesel cetane number from NIR spectra acquired on the hydrocracking total effluent. Red dotted lines: upper and lower limits of the reproducibility of the reference method (±3.6)*

Further analysis of the out-of-limit predicted sample revealed that, compared to the rest of the total effluent samples analyzed, this sample with atypical behavior was obtained under particular process conditions, being the only sample obtained when processing a high paraffinic carbon content feedstock (>60%) at low operating pressure (50 bar). The poor prediction observed raises the question of using this sample in the test data set due to the particularity of its origin. The sample distribution in the data sets was done by applying the unsupervised Kennard-Stone algorithm on the Mx matrix. Therefore, it could be assumed that either the NIR spectrum of this sample does not provide enough detail for the splitting algorithm to classify it in the calibration data set or that the algorithm could have constraints in identifying the specific behavior of the sample. This question can be discussed in future work. However, regardless of the poor sample estimation reason (misclassification of the sample in the test dataset or operating conditions particularity), it is important to address the model robustness constraint to ensure reliable performance over time and under different analytical conditions. In any case, the PLS model developed has a satisfactory performance in estimating the studied property at controlled analysis conditions, either the process or the spectra acquisition.

# 2. Middle distillates properties estimation

The modelling work described in section 1 of this chapter was replicated for the remaining six middle distillate properties. From this work, it can be emphasized that the comparative analysis of the regression methods produced similar results for all six properties. Therefore, the regression method used to calibrate the final models was the PLS. Table 19 summarizes the performance of each developed model regarding errors and squared correlation coefficients, showing the reproducibility of the reference method used to measure the studied properties. The detailed results of the models, including the preprocessing scheme used, are reported in Appendix 6.

*Table 19. NIR models summary for middle distillates properties estimation*

| | Diesel | | | | Kerosene | | |
|---|---|---|---|---|---|---|---|
| | **CN** | **PP** | **CP** | **CFPP** | **CN** | **SP** | **FP** |
| RRM | ±3.6 | ±6.0 | ±4.0 | ±3-0.06*CFPP | ±3.6 | ±3.0 | ±0.071*FP |
| PLS LVs | 9 | 3 | 4 | 9 | 9 | 4 | 8 |
| RMSEC | 1.3 | 4.2 | 3.8 | 3.2 | 0.7 | 1.5 | 1.3 |
| RMSECV | 2.2 | 6.6 | 4.4 | 4.3 | 1.0 | 1.6 | 2.1 |
| RMSEP | 2.0 | 5.6 | 3.0 | 3.9 | 0.6 | 1.5 | 1.9 |
| Bias | -0.6 | -0.5 | -0.2 | -0.1 | 0.2 | -0.2 | 0.4 |
| SEP | 1.9 | 5.6 | 3.0 | 3.9 | 0.6 | 2.5 | 1.9 |
| $r^2C$ | 0.986 | 0.840 | 0.760 | 0.820 | 0.986 | 0.867 | 0.836 |
| $r^2CV$ | 0.959 | 0.598 | 0.678 | 0.687 | 0.970 | 0.843 | 0.613 |
| $r^2P$ | 0.955 | 0.692 | 0.758 | 0.711 | 0.964 | 0.838 | 0.659 |
| RMSE(CV/C) | 1.7 | 1.6 | 1.2 | 1.3 | 1.4 | 1.1 | 1.6 |
| RMSE(P/C) | 1.5 | 1.4 | 0.8 | 1.2 | 0.9 | 1.0 | 1.5 |
| NCD | 67 | 54 | 76 | 65 | 61 | 55 | 41 |
| NTD | 31 | 29 | 30 | 30 | 29 | 24 | 20 |
| %Eff | 97 | 76 | 80 | 77 | 100 | 96 | 100 |
| Limit_min | 30.3 | -42.0 | -31.0 | -29.0 | 21.5 | 13.3 | 42.0 |
| Limit_max | 69.5 | 0.0 | 0.0 | 3.0 | 46.0 | 34.0 | 58.0 |

*RRM = Reproducibility of the Reference Method*
*CN = Cetane Number*
*PP = Pour Point*
*CP = Cloud Point*
*CFPP = Cold Filter Plugging Point*
*FP = Flash Point*
*SP = Smoke Point*
*NCD = Number of Calibration Data*
*NTD = Number of Test Data*
*%Eff = Effectiveness percentage*

The main finding drawn from the results reported is that the models' prediction errors are close to and even lower than the reproducibility of the reference methods. The results validate the feasibility of using the NIR spectra of the HCK total effluent for predicting the middle distillates properties. Table 19 shows that the models having the best performing are those for estimating the cetane number of diesel and kerosene,

evidenced in both the squared correlation coefficients (>0.95) and the percentage prediction effectiveness (>95%). Similarly, the models for predicting the flash and smoke point of the kerosene cut have a satisfactory performance regarding the errors and the prediction effectiveness (>95%). Nonetheless, the $r^2CV$ and $r^2P$ of the flash point are quite low. These values could be related to the number of data available over the entire range of this property evaluation. The lower range of the data set used (42 < FP < 50) has only 17% of the available information, while the remaining 83% is in the upper range (50 < FP < 58) (see Appendix 6, kerosene results).

Concerning the diesel cold flow properties, the models present prediction errors close to the reproducibility of the reference methods. However, it is observed that these models are still susceptible to improvement as the $r^2CV$ and $r^2P$ are lower than 0.8, and the prediction efficiency is inferior to 80%. Figure 10 shows each model performance when estimating these three properties using the external test data set.

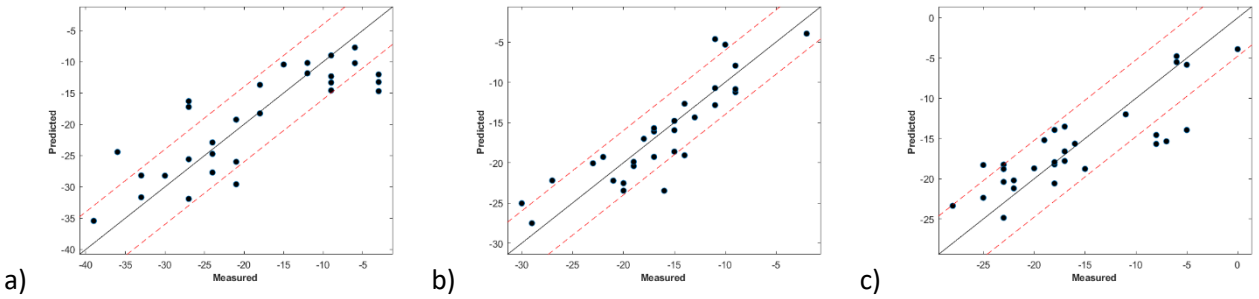



*Figure 10. a) Pour Point, b) Cloud Point, and c) Cold Filter Plugging Point parity plot of NIR PLS models*
*Red dotted lines: upper and lower limits of the reproducibility of the reference method (CP = ±6.0, PP = ±4.0, CFPP = ±3-0.06*CFPP)*

An analysis of the operational information concerning the production of the samples estimated outside the reproducibility limits found no particular or anomalous cause for their poor estimation. The atypical behavior of these samples could be due to the molecular interactions that are not fully captured by the total effluent NIR spectral information. Still, these suboptimal results do not alter the overall finding of this chapter, i.e., the feasibility of predicting middle distillate properties from NIR spectral information acquired on the total effluent of the HCK process.

# 3. Concluding remarks

The results shown in this chapter addressed the first research question by validating the feasibility of estimating middle distillates properties from NIR spectra acquired on the total effluent of the HCK process with errors close to the reproducibility of the reference methods. This first milestone was achieved due to the versatility of the analytical technique NIR in capturing the chemical information of an intermediate process product (total effluent) to reliably describe the behavior of the process final products (middle distillates). This reliable property estimation is possible since diesel and kerosene are embedded in the total

effluent (these two cuts are recovered from the atmospheric distillation of the total effluent). Thus, the investigated characterization alternative offers reliable results since it captures the chemical relationship between these samples (total effluent and middle distillates).

The properties estimation approach investigated can be applied to characterize the different products obtained from the HCK process, even those not included in this thesis (naphtha and unconverted oil). This finding is not limited to the HCK process alone. The results obtained offer the possibility of applying the investigated alternative to processes where a time- and cost-effective product characterization is sought (separation, thermal and catalytic conversion processes), capitalizing on the existing chemical relationship between an intermediate product and the final products.

It is important to consider the analytical technique used and the information it can contribute to the description of the property studied when reproducing the work and the results shown in this chapter. This background knowledge helps establish the chemometric methods best suited to the research objective. For example, the core of this thesis was the analytical technique NIR since its response time is minimal (a few seconds), and it contains chemical information that enables the simultaneous analysis of multiple components. However, the interpretation and exploitation of this technique is not a straightforward task and must be conducted thoroughly by correctly employing the appropriate chemometric methods. One of the steps in chemometric Modelling is the proper selection of preprocessing methods. This step eliminates noisy and non-relevant information, extracting the information that best describes the property under study. During the thesis development, different preprocessing methods were evaluated, demonstrating that the impact of choosing the adequate method is significant. Even with slight variations in the parameters of some preprocessing methods (window-point size in the SavGol method), the model performance is affected. Therefore, it is recommended to perform the preprocessing method selection analysis for each studied property even if the same spectrum is used to estimate several properties. Moreover, it is also suggested to do this analysis when the databases are updated since there is a possibility that the interaction between the existing and added chemical information may be better captured with a different preprocessing method than the one currently used.

Another task to be performed carefully is the generation of datasets for model calibration and testing, especially when large databases (#observations > 500) are not available, as in the case of this thesis. If this task is not implemented carefully, it can lead to flawed conclusions. An example can be found when a constant pattern is observed in developing models where the RMSECV and RMSEP are lower than the RMSEC. Although there are some cases where this trend is plausible, it should be suspected that the sample selection for model calibration and validation is adequate when this trend is the general rule and not the exception. Therefore, regardless of the method used for the database splitting, performing a preliminary analysis of the data (e.g., PCA) to determine the symmetric and coherent distribution of the selected samples in each dataset is highly recommended (see article 2 in appendix 2).

The choice of regression method is equally important to the steps described previously. The most commonly used regression method is PLS, while nonlinear regression methods and the booming use of machine learning methods have been of great interest in recent years. Nonetheless, the latter regression methods' effectiveness is limited when the database size is relatively small. Local regression and neural network methods have a high probability of overfitting problems, affecting model consistency regarding errors and $r^2$'s. The SVM method is an alternative to these methods with a lower risk of model overfitting when a small size database is used.

It shall always be convenient for the researcher to develop interpretable regression models since they enable a deep understanding of the phenomena and parameters affecting the property being studied. One of the disadvantages of machine learning and neural network regression methods, in addition to the complex calibration of their parameters, is the limited interpretability of the generated model. These regression methods are generally employed when a more accurate estimation of the studied property is sought or when it is necessary to consider nonlinearities that linear and conventional methods (PLS) fail to capture. If these methods do not significantly improve model performance over conventional regression methods, it is a good practice to retain interpretable models. Therefore, it is recommended to conduct a comprehensive comparison of the different regression methods available. For example, in this thesis, the regression method used for all properties was PLS since the other evaluated methods did not significantly improve the models' performance. However, by expanding the database with more observations, non-conventional regression methods could potentially give better results, especially when estimating the cold flow properties of diesel.

A last important aspect to consider is the criteria or figures of merit used to select the best regression model. Generally, in a PLS model, the number of latent variables retained is defined as a function of the RMSECV. The method used to calculate this parameter can wrongly influence the decisions made in model calibration. This method should be carefully selected depending on the type and size of the database used. Accordingly, it is recommended to have an external database to complement the analysis of the regression model selection.

In employing the investigated characterization approach, the response time in estimating the studied properties is significantly reduced as the distillation of the total effluent to recover the physical cuts is not required, offering the possibility of performing the properties estimation in real-time. However, it should be stressed that the results and findings discussed in this chapter are valid as long as it is ensured that the NIR spectrum acquisition conditions are repeatable and reproducible regarding those used in the database generation for model calibration. Hence, the performance of the models can be affected by external parameters that impact the quality of the acquired spectrum. Moreover, the obtained results evidenced that the process operating conditions influence the properties estimation. This issue, related to the calibration robustness[157], could be addressed by developing predictive models that simultaneously use the information of the total effluent NIR spectra and the operating conditions employed in obtaining the sample.

In summary, all the studied properties can be accurately estimated from the NIR spectra acquired on the HCK total effluent with errors close to the reference methods. However, the diesel cold flow properties estimation is still susceptible to optimization and improvement to increase its accuracy. Furthermore, considering that the value of these properties can be impacted by different factors such as the type and interaction of the molecules present in the samples, as well as the operating conditions used in the HCK process, it could be expected that the performance improvement of the predictive models could be achieved by using complementary information to the NIR spectra.

# CHAPTER IV
# Data Fusion Modelling

# Chapter IV. Data Fusion Modelling

The seven middle distillates properties studied in this thesis were successfully estimated by PLS models calibrated using NIR spectra acquired on the total effluent of the HCK process. However, some predictive models, such as those used to estimate the cold flow properties of diesel, evidenced opportunities for improvement to increase their accuracy. Therefore, based on the scientific papers 3 and 4 (see appendices 3 and 4, respectively), this chapter addressed the thesis's second research question: "Including additional and descriptive information to the NIR spectra improves the model performance?"

This chapter is divided into three main sections. The first gives an overview of the strategies employed to improve the model performance. The second section presents the results obtained by applying the strategies described in the first section. Finally, the conclusions of the chapter are presented. This chapter presents only the work done on the Cold Filter Plugging Point (CFPP) of diesel. The results obtained on the other diesel cold flow properties, the diesel cetane number and the smoke point of kerosene are given in appendix 6.

## 1. Methodology for improving model performance

The main objective of this chapter is to validate the feasibility of improving the model performance by using complementary information to the NIR spectra. Data fusion modelling was employed to achieve this goal. The first step implemented in this modelling approach was defining and generating the data blocks used to calibrate the models. Compared to the NIR, the $^{13}C$ NMR spectroscopy gives more detailed information on the molecular interactions and bonds present in the analyzed sample[158–160]. Hence, the data extracted from the NMR analytical technique was used as complementary information to improve the model performance (see paper 3 in appendix 3). In addition, to analyze the process variables' impact on the middle distillates properties, the operating conditions utilized in the pilot units and the characterization of the feedstock and total effluent samples were also employed (see paper 4 in appendix 4). In summary, three blocks of data were employed in the data fusion modelling: (i) NIR spectra, (ii) NMR spectra, and (iii) process variables (PVs). Considering the information available of each block, the corresponding independent x-matrices were constructed (Mx1 = NIR [58 x 2180], Mx2 = NMR [58 x 13926], Mx3 = PVs [58 x 53]). The dependent matrix containing the diesel CFPP measurements was also generated. The process variables used are summarized in tables Table *6*, Table *7*, and Table 8 of Chapter 2.

The second step consisted of two stages. The first involved the preprocessing of each data block. As discussed in the preceding chapter, the proper preprocessing scheme selection is crucial in the models' development. For estimating the diesel CFPP, the NIR data block was preprocessed using the variable sorting for normalization (VSN) method, followed by the Savitzky-Golay third derivative using a 9-point window and a

fourth-order polynomial (SavGol[9,4,3]). Regarding the NMR spectra, they were first aligned using the Interval Correlation Optimized (icoshift) algorithm. Subsequently, the spectra were preprocessed using the Savitzky-Golay smoothing (Smooth SavGol) and normalized (each variable is divided by the sum of the absolute values of all the variables for a given spectrum). Lastly, the block of process variables was autoscaled to prevent their natural scale from influencing the result (see paper 4 appendix 4). The second stage concerned the generation of the calibration and test data sets by applying the Kennard-Stone algorithm on the Mx1.

The third step was the calibration of the regression models. First, three individual models were calibrated using each block separately. Then, evaluating the performance of each fusion strategy summarized in Table 13 of Chapter 2 (see paper 3 in appendix 3), three data fusion model sets were developed: (i) NIR + NMR, (ii) NIR + PV, (ii) NIR + NMR + PV. Finally, these models were analyzed to determine the best-performing model compared to the single NIR model.

Once data fusion was applied, it was decided to evaluate the potential of variable selection to further improve the models' performance. First, the variable selection was applied to the different blocks. Table 14 of Chapter 2 summarizes the methods applied on the PVs data block, while Table 15 shows the methods applied on the multivariate data blocks (NIR, NMR). Next, the best-performing model in each set of data fusion models without variable selection was recalibrated, and the results were compared (see paper 4 in appendix 4).

The schematic diagram shown in Figure 11 summarizes the steps implemented to improve the performance of the NIR models. The detailed description of the methodology employed for selecting the best data fusion strategy and variable selection method for improving the NIR model performance is discussed in appendices 3 and 4.

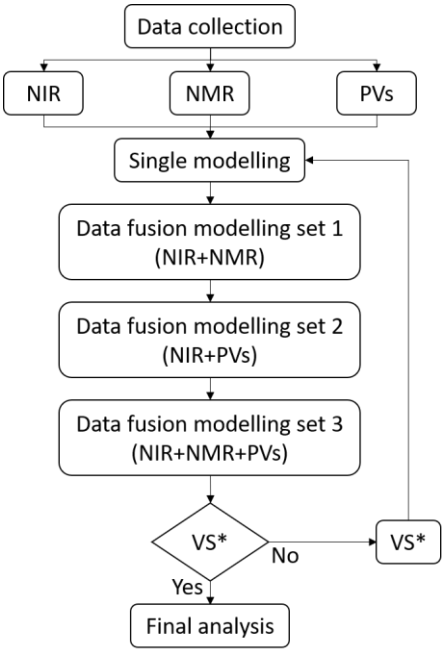



*Figure 11. Flow diagram describing the steps involved in improving the NIR model performance. *VS = Variable Selection applied*

## 2. Results analysis

The single models using the multivariate data blocks (NIR and NMR) without applying variable selection were calibrated using the PLS regression method, while the MLR method was used for calibrating the model from the process variable data (Mx3). Concerning the data fusion models, the best strategy for fusing the Mx1 and Mx2 blocks was the mid-level using the PLS model scores calibrated from each data block as the fusion features (see appendix 3). On the other hand, when fusing the multivariate data blocks and the process variables, the strategy having the best results was the high-level data fusion using the predicted variable for each block as the fusion decision. Table 20 summarizes the results of the models developed using all the variables available in each data block.

*Table 20. Single and data fusion models summary for diesel CFPP estimation using all available variables*

| | **Cold Filter Plugging Point (°C)** | | | | | |
|---|---|---|---|---|---|---|
| Model | **1** | **2** | **3** | **4** | **5** | **6** |
| RRM | ±3-0.06*CFPP | | | | | |
| RMSEC | 3.2 | 3.1 | 3.3 | 1.4 | 1.5 | 1.5 |
| RMSECV | 4.6 | 4.6 | 4.7 | 1.9 | 1.6 | 1.6 |
| RMSEP | 3.6 | 3.1 | 3.6 | 2.2 | 2.3 | 2.2 |
| Bias | -0.9 | 0.3 | 0.7 | 0.5 | -0.3 | 0.3 |
| SEP | 3.5 | 3.1 | 3.5 | 2.1 | 2.3 | 2.2 |
| $r^2C$ | 0.844 | 0.856 | 0.831 | 0.971 | 0.968 | 0.967 |
| $r^2CV$ | 0.693 | 0.685 | 0.670 | 0.945 | 0.962 | 0.963 |
| $r^2P$ | 0.795 | 0.849 | 0.792 | 0.925 | 0.914 | 0.920 |
| RMSE(CV/C) | 1.4 | 1.5 | 1.4 | 1.4 | 1.1 | 1.1 |
| RMSE(P/C) | 1.1 | 1.0 | 1.1 | 1.6 | 1.5 | 1.5 |
| NCD | 40 | | | | | |
| NTD | 18 | | | | | |
| %Eff | 78 | 78 | 72 | 94 | 89 | 94 |
| Limit_min | -29.0 | | | | | |
| Limit_max | 3.0 | | | | | |

*1 = NIR, 2180 variables, PLS 4 LVs*
*2 = NMR, 13926 variables, PLS 4 LVs*
*3 = PVs, 53 variables, MLR*
*4 = NIR [6LVs] + NMR[6LVs], mid-level scores PLS, PLS 6 LVs*
*5 = NIR [6LVs] + PVs[53 var], high-level, predicted CFPP, MLR*
*6 = NIR [6LVs] + NMR[6LVs] + PVs[53 var], high-level, predicted CFPP, MLR*
*RRM = Reproducibility of the Reference Method*
*NCD = Number of Calibration Data*
*NTD = Number of Test Data*
*%Eff = Effectiveness percentage in predicting new samples*

The performance of the three individual models regarding the RMSEC and RMSECV is comparable. However, concerning the other statistical parameters, the single model with the lowest prediction bias, lowest RMSEP, and highest $r^2P$ is the one calibrated from NMR spectra. These results show a better capture of the

relationship between the chemical information of the total effluent and the studied property. Furthermore, although the percentage of effectiveness in predicting new samples within the reference method reproducibility limits is the same as the NIR model, the NMR model presents a higher accuracy over the whole range of property evaluation (see Figure 12a,b). In turn, the performance of the model calibrated from PVs is significantly similar to the NIR model.

From Table 20, it is observed that by the synergic use of the complementary information to the NIR block, the performance of the predictive models improves significantly. Compared to the single NIR model, the data fusion models achieve average reductions in RMSEC and RMSEP of 54% and 38%, respectively. The RMSECV has the largest reduction (≈63%). The model performance improvement is also reflected in the prediction bias reduction (≈60%) and the squared correlation coefficients by showing an increase of 15%, 38%, and 16% for $r^2C$, $r^2CV$, and $r^2P$, respectively. The increasing model accuracy over the entire range of model application (see Figure 12) is another advantage observed when data fusion is employed, resulting in higher prediction effectiveness. However, 100% of effectiveness is still not achieved.

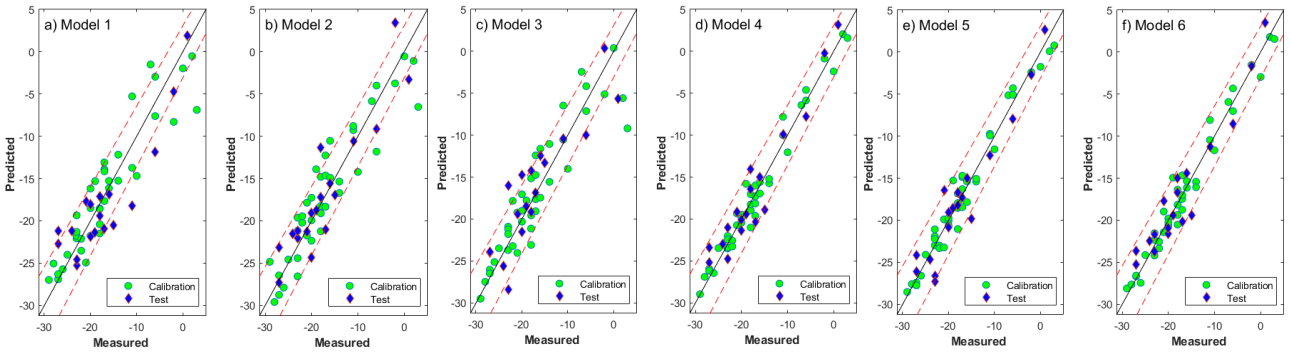



*Figure 12. Parity plot for model performance comparison with no variable selection. a) PLS model from Mx1, b) PLS model Mx2, c) MLR model from Mx3, d) data fusion model using Mx1+Mx2, e) data fusion model using Mx1+Mx3, f) data fusion model using Mx1+Mx2+Mx3. Red dotted lines: upper and lower limits of the reproducibility of the reference method*

The results previously discussed validate the viability of enhancing model performance when complementary information to the NIR spectrum is employed. However, some of the information may be redundant or non-descriptive enough to improve the property estimation. Therefore, the variable selection was applied to each data block to determine if identifying relevant descriptors leads to further improvement in model performance.

From the various methods evaluated for selecting variables in the multivariate data blocks, the best performing was CovSel. Regarding the process variables, the Backward-SFFS method gave the best results in identifying the appropriate descriptors (see paper 4 in appendix 4). After applying the variable selection in each data block, the calibration of individual and data fusion models was repeated. For this work step (variable selection), all individual models were developed using the MLR regression method. In addition, the high-level strategy was employed for the data fusion models using the CFPP predicted by each block as the fusion decision. Table 21 summarizes the results of the models developed using the variables selected in each

data block.

Compared to the independent NIR model using all variables (Table 20 model 1), the RMSECV and RMSEP were reduced by about 17% when the single model was calibrated after applying variable selection on this block (Table 21 model 7). Although the prediction effectiveness remained unchanged, the accuracy over the entire range of model application is evident (see Figure 12a vs. Figure 13a). When the variable selection was applied, the model calibrated from the process variables showed the greatest reduction in prediction bias (57%) and RMSEP (42%). The error reduction of this model was reflected in the prediction accuracy improvement, the $r^2P$ and the percentage of effectiveness (see Figure 12c vs. Figure 13c). Lastly, it can be highlighted the higher effectiveness in predicting new samples achieved in the NMR model when variable selection is applied. Nonetheless, the RMSEC and $r^2C$ of this model were negatively impacted. This punctual model deterioration is due to the limited description of the calibration samples when the diesel CFPP is higher than -5°C. Analyzing Figure 13b, it is observed that most of the calibration data (82%) is in the range where the CFPP is lower than -5°C, which could affect the selection of variables in this data block omitting the information that best describes the samples with CFPP higher than -5°C. Regardless, it is evident that the variable selection helps better estimate the studied property than using all the available variables.

*Table 21. Single and data fusion models summary for diesel CFPP estimation using selected variables*

| | **Cold Filter Plugging Point (°C)** | | | | | |
|---|---|---|---|---|---|---|
| Model | **7** | **8** | **9** | **10** | **11** | **12** |
| RRM | ±3-0.06*CFPP | | | | | |
| RMSEC | 3.2 | 3.7 | 3.2 | 1.8 | 1.4 | 1.2 |
| RMSECV | 3.8 | 4.5 | 3.8 | 2.1 | 1.6 | 1.3 |
| RMSEP | 3.0 | 3.1 | 2.1 | 2.1 | 2.0 | 1.8 |
| Bias | -0.8 | 0.6 | 0.3 | 0.6 | -0.3 | 0.0 |
| SEP | 2.9 | 3.0 | 2.1 | 2.0 | 2.0 | 1.8 |
| $r^2C$ | 0.839 | 0.787 | 0.842 | 0.953 | 0.967 | 0.981 |
| $r^2CV$ | 0.778 | 0.702 | 0.774 | 0.935 | 0.964 | 0.976 |
| $r^2P$ | 0.860 | 0.848 | 0.927 | 0.932 | 0.934 | 0.949 |
| RMSE(CV/C) | 1.2 | 1.2 | 1.2 | 1.2 | 1.1 | 1.1 |
| RMSE(P/C) | 0.9 | 0.8 | 0.7 | 1.2 | 1.4 | 1.5 |
| NCD | 40 | | | | | |
| NTD | 18 | | | | | |
| %Eff | 72 | 83 | 94 | 100 | 100 | 100 |
| Limit_min | -29.0 | | | | | |
| Limit_max | 3.0 | | | | | |

*7 = NIR, 5 variables, MLR*
*8 = NMR, 4 variables, MLR*
*9 = PVs, 11 variables, MLR*
*10 = NIR [5 var] + NMR[4 var], high-level , MLR*
*11 = NIR [5 var] + PVs[11 var], high-level, predicted CFPP, MLR*
*12 = NIR [5 var] + NMR[4 var] + PVs[11 var], high-level, predicted CFPP, MLR*
*RRM = Reproducibility of the Reference Method*
*NCD = Number of Calibration Data*

*NTD = Number of Test Data*
*%Eff = Effectiveness percentage in predicting new samples*

When applying data fusion using only the variables identified and selected in each data block, it was possible to achieve 100% effectiveness in predicting new samples within the reproducibility limits. Compared to the single NIR models, the improvement in the diesel CFPP estimation is observed in all the statistical parameters calculated. The highest performance in predicting the property studied is obtained when the three data blocks are fused (model 12). In this data fusion model, the $r^2$'s are equal to or greater than 0.95, the prediction bias is negligible, and all samples, both calibration and test, are predicted within the reproducibility limits (see Figure 13f). The second best-performing model is the data fusion model between the NIR and PV blocks (model 11). An advantage of this latter model over the preceding one is the possibility to be applied in real-time.

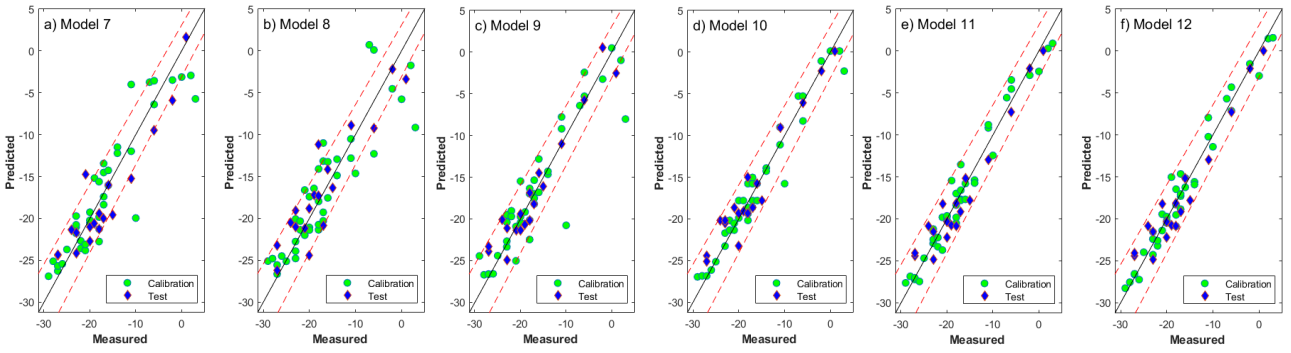



*Figure 13. Parity plot for model performance comparison with variable selection. a) MLR model from Mx1, b) MLR model Mx2, c) MLR model from Mx3, d) data fusion model using Mx1+Mx2, e) data fusion model using Mx1+Mx3, f) data fusion model using Mx1+Mx2+Mx3. Red dotted lines: upper and lower limits of the reproducibility of the reference method*

Figure 14 compares all the models developed (individual and data fusion, with and without variable selection) to perform an integrated analysis in graphical form. From this figure, the two main conclusions of the work shown in this chapter can be corroborated. First, the synergic use of complementary information to the NIR spectroscopy improves the estimation of the diesel CFPP. Second, the estimation can be further enhanced by applying variable selection on the data blocks prior to their fusion.

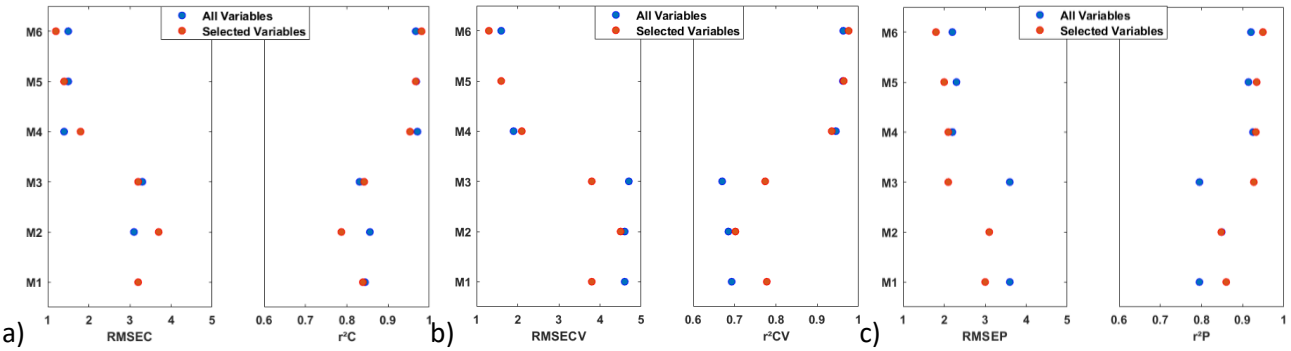



*Figure 14. Model performance comparison a) RMSEC & $r^2$C. b) RMSECV & $r^2$CV. c) RMSEP & $r^2$P*
*Models using data blocks. M1: NIR, M2: NMR, M3 :PVs, M4: NIR+NMR, M5: NIR+PVs, M6: NIR + NMR + PVs*

# 3. Concluding remarks

The analysis presented in this chapter addressed the second research question related to the problem of improving the prediction model performance using supplementary information to NIR spectroscopy. The results demonstrated that the information from different sources, either analytical or operational, can be used synergically to improve the studied property description. One of the model performance improvements achieved by data fusion is the prediction bias reduction, leading to a more accurate property prediction. This improvement is due to the supplementary information provided by each data block. The NMR data block provides a more detailed description of the total effluent's chemical composition, helping to explain better the chemical relationship of this sample with the middle distillates. In turn, the block of process variables contributes to the property estimation improvement by describing the impact of these variables on the properties studied.

The results have validated the performance improvement of the models by data fusion modelling. Nevertheless, it is imperative to consider the constraints linked to their implementation. For example, if it is planned to use these models in online monitoring applications, it must be considered which information can be acquired in real-time. In this case, models using NMR spectra and total effluent properties cannot be used to estimate middle distillate properties in real-time. Although it is possible to acquire this information online, this alternative would imply additional data acquisition costs. Therefore, it is recommended to analyze the cost-benefit ratio of using this information to improve the accuracy of property estimation. Alternatively, if it is intended to improve model performance while keeping data acquisition costs low, data fusion models between NIR spectra, operating conditions and feedstock properties are a suitable alternative. While the performance of these models is lower than those using NMR spectra and total effluent properties, compared to the base NIR model, the property estimation is more accurate.

The proper choice of the strategy and data fusion methods is essential to ensure the most optimal model calibration. While some results may lead to a general decision to choose or discard a strategy or data fusion method, it is strongly recommended to evaluate the effectiveness of each method based on the research objective. This thesis found that certain data fusion strategies and methods have limited performance when fusing highly multivariate data blocks (NIR and NMR) with low multivariate data blocks (process variables). Namely, the low-level data fusion strategy, including the SO-PLS and ROSA methods. These findings validate the recommendation formerly suggested while enabling the development of new strategies to achieve a suitable fusion between these heterogeneous data. For example, the different data fusion levels (low-, mid-, high-) are normally used independently. However, for one of the diesel cold flow properties (cloud point), it was found that the combination of two fusion levels offered the best result. First, the NIR and NMR data blocks were fused using the mid-level data fusion employing the scores of PLS models developed from each

block as the fusion features. Then, using the high-level fusion, the values predicted by the model developed from the process variables and the values predicted by the data fusion model of the spectroscopic blocks were used to calibrate the final model. These results showed that different strategies could be evaluated and applied to obtain an optimal performance model.

In addition to choosing the most appropriate strategy and method for data fusion, the proper configuration of some data fusion methods is also important. An example can be found in using the SO-PLS method. The performance of this method is influenced by the order in which the data blocks are used. For properties where a more precise molecular description is needed, such as the diesel cold flow properties, the best results employing this data fusion method were obtained when the first block analyzed was the NMR. On the contrary, the best results were observed when the NIR block was first analyzed for the other middle distillate properties. For this reason, it is recommended to evaluate different use settings on the same method, in this case, SO-PLS.

Similar to developing individual models (using a single block of data), splitting the database into calibration and test sets is key to obtaining representative and consistent models. Throughout the development of the thesis, it became evident that the data fusion results were impacted when the dataset generation was conducted by applying the Kennard-stone algorithm on a specific data block (either NIR, NMR or process variables). Different evaluations found that in most cases, the best results were obtained when the data splitting algorithm was applied to the NIR block.

The data fusion approach investigated in this chapter has potential use for fast and accurate properties prediction where the performance of single models is limited compared with the reproducibility of the reference method. As discussed previously in this chapter, the data fusion modelling using all variables from each data block improves the estimation of the diesel CFPP compared to single models. However, applying variable selection to each data block before data fusion significantly improves the estimation of this property and leads to greater model consistency regarding the RMSE's and $r^{2}$'s.

It is important to point out that the choice of the variable selection method must be made carefully according to the characteristics of the analyzed information. The performance of a variable selection method on one data block is not necessarily reflected when applied to another block of data. Indeed, some variable selection methods are very effective in identifying relevant descriptors in multivariate data blocks, but when applied to a less multivariate data block (process variables), their performance is limited, as is the case of the CovSel method. Hence, it is important to use the appropriate variable selection method.

It should be kept in mind that variable selection does not have the sole objective of increasing the performance of the developed models. Depending on the researcher's objective, the efficiency of variable selection methods can be evaluated under completely different criteria. For example, if the main objective is to improve the accuracy of property prediction, the application of variable selection on each data block is the most advisable approach since it extracts from each block the variables that best describe the studied

property. On the contrary, if the objective is to reduce costs regarding the sensors used to capture the information, a multi-block variable selection method, such as SO-CovSel, appears as one of the best methods, even when having heterogeneity in the data blocks (see paper 4 in appendix 4). The results obtained in this thesis showed that a maximum reduction of variables needed in each block is achieved using this method while preserving the performance of the single model. This effectiveness in selecting the pertinent variables is because this method efficiently captures the interaction between the information of each block, removing the redundancy that may exist between them.

The scheme followed in this chapter to improve the models' performance consisted of first performing the variable selection, followed by the data fusion modelling. During the discussion of the results, the question arose as to whether the scheme order could be reversed. For example, performing a mid-level data fusion by concatenating the multivariate PLS model scores for a subsequent variable selection (scores of each model). Although it could result in something unreasonable or not practical, the versatility offered by these methods permits to suggest and apply new strategies when using them.

The methodology for model performance improvement described in the previous sections was applied to the other middle distillate properties, obtaining results comparable to those presented in this chapter (see appendix 6). As mentioned in chapter 3, the measured values of the studied properties can be influenced by factors including the molecular behavior of the samples and the process operating conditions. While it has been shown in the present chapter that using supplementary information has the potential to increase the model accuracy by capturing the influence that these factors have on the properties analyzed, the model performance could still be affected by external parameters associated with the data acquisition conditions, particularly to the NIR spectra. Therefore, it is necessary to ensure the robustness of the model for an accurate and long-term reliable prediction.

# CHAPTER V
# Robust Modelling

# Chapter V. Robust Modelling

The middle distillates properties studied in this thesis were estimated through PLS regression models calibrated from the NIR spectra acquired on the total effluent of the HCK process with errors close to or even lower than the reproducibility of the reference methods. The accuracy in estimating some properties (diesel cold flow properties, diesel cetane number and kerosene smoke point) was improved by using complementary information to the NIR spectra. Except for models calibrated using NMR spectra and total effluent properties as complementary information, all developed NIR models have the potential to be used in real-time property estimation applications. However, the limited robustness of the models, defined as the ability to maintain a reliable performance under different application conditions [157, 161], may affect their performance due to the different external parameters associated with the acquisition of the NIR spectra. Based on the results discussed in the scientific paper 5 (see appendix 5), this chapter presents the work conducted to address the third research question formulated: "Can external parameters' influence on the NIR spectra quality be compensated/corrected to ensure a reliable properties estimation over time?". The external parameters investigated were classified into three main groups: (i) instrumental disturbances, (ii) sample temperature, and (iii) factors associated with dynamic spectra acquisition.

This chapter is divided into three main sections. The first section summarizes the strategy employed for correcting the impact of external parameters. The second section shows the results obtained by applying the approaches described in the first section, while the third section summarizes the main conclusions. In this chapter, only the work results related to the diesel cetane number are presented. The results of the other middle distillate properties are summarized in appendix 6.

## 1. Methodology for correcting external parameters impact

To address the issue of model robustness, the general strategy proposed by Chauchard et al. [149] for correcting external parameters impact was employed. The development of each step involved in this general strategy is briefly described hereafter, along with the information used.

Step 1 consisted of defining the stable acquisition conditions, i.e., with no influence of any external parameter, to calibrate the reference PLS model. To be consistent with the progress of the thesis, the conditions used to develop the PLS model for estimating the diesel cetane number described in Chapter 3 were adopted as stable conditions. In summary, 98 spectra were acquired on 98 total effluent samples at a sample temperature of 60°C, in steady-state conditions, using a reflectance probe with an optical length of 2mm. 67 of these 98 samples were used for model development, and the remaining 31 were used for model testing.

Step 2 included establishing the external parameters G and their respective g-values to be studied and performing the spectra acquisition at the defined evaluation conditions. As outlined in the chapter introduction, three general external parameters were evaluated in this thesis. First, the change of the optical length (from 2mm to 1mm) was evaluated for instrumental disturbances. For this parameter, 27 spectra were acquired on 27 samples under steady-state conditions using the same reflectance probe utilized in step 1 but with an optical length of 1mm. Regarding the sample temperature, four temperature levels (60°C - 90°C ΔT = 10°C) were evaluated using the same reflectance probe already discussed with an optical length of 2 mm at steady-state conditions. As a result, 108 spectra were acquired on the 27 previously outlined samples to evaluate the impact of sample temperature. Finally, the dynamic acquisition of NIR spectra was conducted on four different samples for 30 minutes, increasing the sample temperature by 10°C every 10 minutes and using a transmittance flow cell with an optical length of 1 mm. In the dynamic acquisition, 211 spectra were obtained. Details of the external parameters evaluated and the acquisition of the spectra at the defined conditions are reported in scientific paper 5 (appendix 5).

The third step involved determining whether each external parameter significantly impacted the reference model performance developed in step 1. Therefore, the PLS model calibrated at steady-state conditions was tested using the spectra acquired at the previously described conditions. To accomplish this step, 286 spectra were utilized (76 acquired at steady conditions and 210 at dynamic conditions) (see appendix 5).

The fourth step concerned the evaluation of four approaches for correcting the impact of external parameters investigated:

(i) Transfer function development. For this approach, the Piecewise Direct Standardization (PDS) method was used to develop a transfer function to be applied to the NIR spectra acquired at different conditions prior to their use in testing the reference PLS model developed at steady conditions. In this approach, two categories of transfer functions were developed. The first one with a global function enabling the simultaneous correction of all the studied parameters' impact. The second one with a transfer function developed for each parameter to be corrected, i.e., two transfer functions in total (instrumental disturbances & sample temperature). A transfer function for the dynamic acquisition was not possible since no reference spectra of the samples used in this type of acquisition were available. Therefore, the cetane number estimation at dynamic conditions was performed using the appropriate transfer function (optical length or sample temperature transfer functions).

(ii) Calibration of an individual PLS regression model for each external parameter evaluated. Five different PLS models were developed. One at steady-state conditions, one for spectra acquired with an optical length of 1 mm, and one PLS model for each sample temperature other than 60°C evaluated (3 models in total - 70°C, 80°C, 90°C). The most suitable developed model was used according to the acquired acquisition conditions for estimating the diesel cetane number using the spectra acquired at dynamic conditions (optical length and sample temperature).

(iii) Global PLS model development using spectra obtained at different acquisition conditions in the model calibration dataset. In this approach, the Orthogonal Signal Correction (OSC) preprocessing method using nine components was complementary applied.

(iv) Robust modelling using orthogonalization methods. In this last approach, the External Parameter Orthogonalization (EPO) and Dynamic Orthogonal Projection (DOP) methods were used to develop a robust model that simultaneously corrected the different external parameters evaluated. A detailed description of the robust model development can be found in paper 5 (appendix 5).

## 2. Results analysis

As mentioned in section 1 of this chapter, the reference PLS model used to evaluate the impact of the different external parameters was the one described in chapter 3. As a reminder, this model employs 9 latent variables having an RMSEC and RMSECV of 1.3 and 2.2, respectively. Table 22 column 1, summarizes the performance of this model as a function of SEP, bias, RMSEP and $r^2P$ using the spectra acquired at different measurement conditions. In Table 22, these assessment metrics have been calculated separately for each external parameter investigated and globally.

From the results reported, it is possible to conclude that as long as the model is applied to NIR spectra acquired at the previously described steady-state conditions, its performance is satisfactory. However, it is observed that the different external parameters studied significantly affect the estimation of the diesel cetane number. This conclusion can be corroborated in the parity plot shown in Figure 15a. Excluding the samples with the highest and lowest cetane number values, the change in optical length is the parameter with the least impact, partially correctable by preprocessing the new spectra. On the other hand, the spectra acquired at different sample temperatures generate predictions with a low SEP (2.8) but a high bias (-12.2), resulting in a high RMSEP (12.5). The preprocessing scheme employed in developing the reference model fails to correct such bias. The analysis aforementioned is reflected in the high $r^2P$ value (0.938) of the samples predicted at different temperatures. Finally, the factors associated with the spectra dynamic acquisition have the greatest impact on the model performance resulting in a moderate value of SEP (3.3), with high values of bias (-27.4) and RMSEP (27.6), and a low $r^2P$ (0.065). When evaluating all parameters simultaneously, it is found that only 14% of the samples used to test the model are predicted within the reproducibility limits of the method. This effectiveness percentage corresponds mainly to the predicted samples whose NIR spectra were acquired at reference conditions.

*Table 22. Approaches effectiveness comparison for correcting external parameters impact on diesel cetane estimation*

| | Cetane Number | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | **1** | **(i) 2** | **(i) 3** | **(ii) 4** | | | | | **(iii) 5** | **(iv) 6** |
| RMSEC | 1.3 | 1.3 | 1.3 | 1.3 | 1.6 | 1.2 | 1.3 | 1.7 | 1.6 | 1.6 |
| RMSECV | 2.2 | 2.2 | 2.2 | 2.2 | 2.6 | 2.2 | 2.3 | 2.4 | 2.2 | 2 |
| $r^2C$ | 0.986 | 0.986 | 0.986 | 0.986 | 0.981 | 0.988 | 0.986 | 0.977 | 0.979 | 0.981 |
| $r^2CV$ | 0.959 | 0.959 | 0.959 | 0.959 | 0.946 | 0.965 | 0.960 | 0.956 | 0.965 | 0.969 |
| RMSEP_P0 | 2.0 | 2.0 | 2.0 | 2.0 | | | | | 1.9 | 1.8 |
| Bias_P0 | -1.0 | -1.0 | -1.0 | -1.0 | | | | | -0.1 | -0.2 |
| SEP_P0 | 1.7 | 1.7 | 1.7 | 1.7 | | | | | 1.9 | 1.8 |
| $r^2P$_P0 | 0.971 | 0.971 | 0.971 | 1.0 | | | | | 0.968 | 0.968 |
| RMSEP_P1 | 10.1 | 4.5 | 6.2 | 1.9 | | | | | 2.3 | 2.3 |
| Bias_P1 | -7.9 | -1.6 | -3.0 | -0.1 | | | | | 1.2 | 0.9 |
| SEP_P1 | 6.3 | 4.2 | 5.5 | 1.9 | | | | | 2.0 | 2.1 |
| $r^2P$_P1 | 0.680 | 0.870 | 0.810 | 1.0 | | | | | 0.968 | 0.972 |
| RMSEP_P2 | 12.5 | 3.9 | 3.2 | 1.7 | | | | | 2.3 | 2.5 |
| Bias_P2 | -12.2 | -3.0 | -2.2 | -0.2 | | | | | 1.0 | 0.8 |
| SEP_P2 | 2.8 | 2.6 | 2.3 | 1.7 | | | | | 2.1 | 2.4 |
| $r^2P$_P2 | 0.938 | 0.959 | 0.961 | 1.0 | | | | | 0.965 | 0.954 |
| RMSEP_P3 | 27.6 | 12.6 | 13.1 | 25.1 | | | | | 3.2 | 2.1 |
| Bias_P3 | -27.4 | -12.3 | -12.8 | 24.9 | | | | | -1.5 | 0.1 |
| SEP_P3 | 3.3 | 2.8 | 2.8 | 2.9 | | | | | 2.9 | 2.1 |
| $r^2P$_P3 | 0.065 | 0.036 | 0.069 | 0.068 | | | | | 0.071 | 0.366 |
| RMSEP_PG | 24.1 | 10.9 | 11.4 | 21.5 | | | | | 3.0 | 2.1 |
| Bias_PG | -21.7 | -9.5 | -9.9 | 18.1 | | | | | -1.0 | 0.1 |
| SEP_PG | 10.4 | 5.4 | 5.7 | 11.6 | | | | | 2.8 | 2.1 |
| $r^2P$_PG | 0.019 | 0.456 | 0.410 | 0.7 | | | | | 0.850 | 0.917 |
| NTD | 287 | | | | | | | | | |
| %Eff_PG | 14 | 22 | 23 | 25 | | | | | 85 | 92 |
| RRM | ±3.6 | | | | | | | | | |

*1 = Reference PLS NIR Model 9 LVs*
*(i) 2 = General transfer function using PDS (11 point-window)*
*(i) 3 = Specific transfer function using PDS (11 point-window) (Optical length and Sample T)*
*(ii) 4 = Specific model for each external parameter (See Table 34)*
*(iii) 5 = General PLS NIR model with 9 LVs using spectra acquired at different conditions*
*(iv) 6 = PLS_EPO_DOP model with 8 LVs and 12 EPO components*
*P0 = No external parameters evaluated (Sample T = 60°C, probe optical length = 2mm)*
*P1 = Optical length impact evaluation (Sample T = 60°C, probe optical length = 1mm)*
*P2 = Sample T impact evaluation (Sample T = 60°C - 90°C, probe optical length = 2mm)*
*P3 = Dynamic impact evaluation (Sample T = 60°C - 90°C, flowcell optical length = 1mm)*
*PG = All external parameters evaluated simultaneously*
*RRM = Reproducibility of the Reference Method*
*NTD = Number of Test Data*
*%Eff = Effectiveness percentage in predicting new samples*

When applying the first correction approach (transfer function using the PDS method), it can be observed that the bias caused by the different external parameters is reduced, improving the RMSEP as well. Compared to the reference PLS model, the best-corrected external parameter is the optical length change with an

average bias reduction of 79%, followed by the bias caused by the sample temperature, which is reduced by about 76%, while the bias caused by the dynamic acquisition is reduced by 55%. In addition to the observed improvement in errors and $r^2$ values, the percentage of predicted samples within reproducibility limits increases by 8 percentage points. Despite the improvements achieved with this correction approach, the model still does not perform optimally, having an overall RMSEP (10.9) higher than the reproducibility of the reference method (±3.6). The factors associated with the dynamic acquisition of the spectrum are the ones that contribute the most to this error value. Completing the analysis of this first approach, it can be highlighted that a greater correction of the overall parameters impact is achieved when a general transfer function is developed to correct simultaneously the parameters' influence, rather than when a transfer function is developed for each external parameter (see Figure 15b and c).

The second approach applied (development of a PLS model for each g value of the G parameters) showed promising results in correcting the impact caused by the change of optical length and sample temperature by reducing the bias by approximately 98%, resulting in RMSEP values (≈1.8) lower than the reproducibility of the reference method. However, this approach presents two main drawbacks. The first one is related to the low performance of the models developed to estimate the diesel cetane number using the spectra acquired under dynamic conditions. Table 22 column 4 shows that, compared to the reference PLS model, the bias has a very low reduction (9%). This lack of robustness to predict the diesel cetane number under dynamic conditions results in the global RMSEP being very similar to the obtained by the reference model without applying any correction (see Figure 15d). The second drawback of the evaluated approach concerns the complexity of applying multiple preprocessing methods and models to predict the same property (see Table 23).

*Table 23. Individual models description for correcting external parameters impact on diesel cetane estimation*

| Model | Description | Preprocessing scheme | Latent Variables |
|---|---|---|---|
| (ii) 4a | To estimate at steady conditions | SNV + SavGol [23,3,2] | 9 |
| (ii) 4b | To correct optical length change | MSC + SavGol [23,2,2] | 5 |
| (ii) 4c | To correct sample temperature (70°C) | SNV + SavGol [17,2,2] | 6 |
| (ii) 4d | To correct sample temperature (80°C) | MSC + SavGol [23,3,2] | 6 |
| (ii) 4e | To correct sample temperature (90°C) | PQN + SavGol [25,2,2] | 5 |

The third approach employed (development of a global model using spectra acquired at different measurement conditions) succeeded in overcoming the two limitations exhibited by the second approach. With an average reduction of 90% of the bias caused by each external parameter, including dynamic acquisition, this approach achieves a percentage of predicted sample effectiveness of 85% (see Figure 15e). This effectiveness improvement is reflected in the reduction of bias (96%), SEP (73%), and hence RMSEP (88%) when evaluating the impact of external parameters simultaneously. In addition, the $r^2P$ is also significantly improved (98%). In summary, using this approach, the correction of the impact caused by the evaluated external parameters is considerable, given the advantage of using a single model. However, some

estimations at dynamic conditions are still predicted outside the reproducibility limits.

By applying the fourth correction approach (robust modelling with orthogonal methods), the best results are obtained regarding the impact correction caused by the G-parameters. With a single regression model, a prediction effectiveness percentage of 92% (see Figure 15f), a reduction of bias, SEP, and overall RMSEP of 99%, 80%, and 91%, respectively, are achieved. Similarly, the highest $r^2P$ improvement (99%) is achieved with the robust approach adopted. An advantage of using this approach is that there is no need for further transformation of the new spectra being analyzed.

Based on the results obtained and analyzed in this chapter, the feasibility of achieving reliable estimates of the studied property under different application conditions can be validated, being the robust modelling using orthogonalization methods the approach that showed the best performance (see appendix 6).

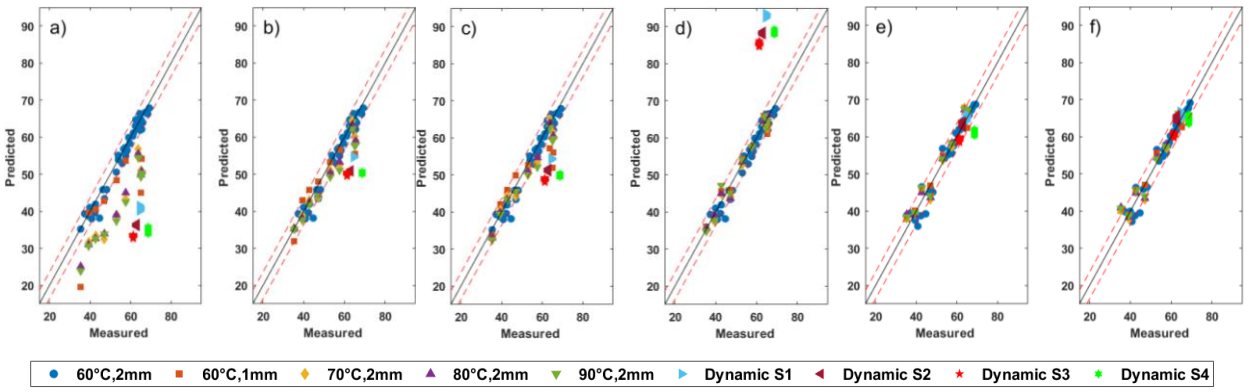



*Figure 15. Parity plot for model performance comparison applying external parameters correction. a) reference PLS model b) approach (i) global PDS transfer function, c) approach (i) individual PDS transfer function, d) approach (ii) individual modelling, e) approach (iii) global modelling, f) approach (iv) robust modelling using orthogonalization. Red dotted lines: upper and lower limits of the reproducibility of the reference method*

At last, Figure 16 compares all the correction approaches implemented regarding the RMSE's and $r^2$'s. The analysis conducted and the main conclusion drawn previously can be corroborated from this figure. Namely, the performance of the NIR model for estimating the diesel cetane number is affected by external parameters associated with the spectra acquisition conditions. The approach that best corrects this impact is robust modelling using the EPO and DOP methods synergically (see paper 5 in appendix 5).

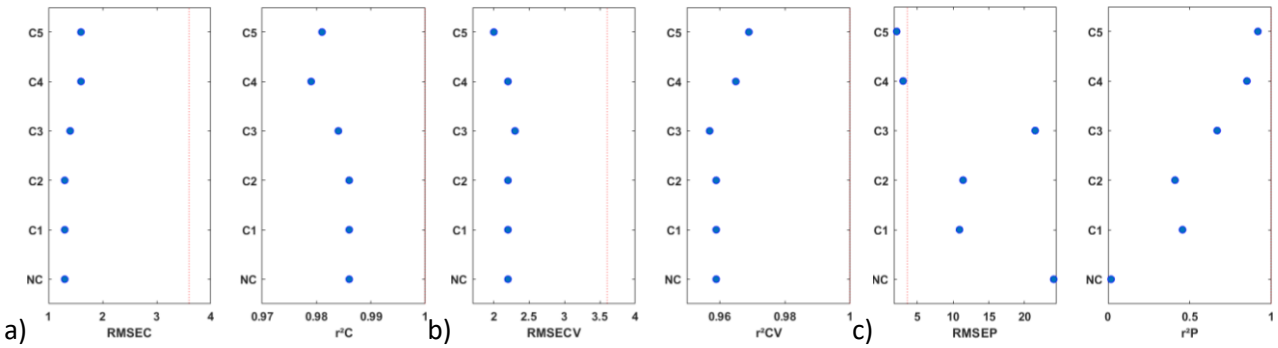



*Figure 16. Model performance comparison applying external parameters correction a) RMSEC & $r^2C$. b) RMSECV & $r^2CV$. c) RMSEP & $r^2P$. Legend. NC = no correction, C1 = correction using global PDS transfer function, C2 = correction using individual PDS transfer function, C3 = correction using individual modelling, C4 = correction using global modelling, C5 = correction using robust modelling*

# 3. Concluding remarks

The analysis presented in this chapter addressed the third research question related to the NIR model robustness issue. The results obtained validated that reliable estimation of diesel cetane number at different acquisition conditions of the NIR spectra is feasible. In this thesis, three general external parameters were studied. Among them, the sample temperature and factors associated with the dynamic acquisition are the parameters that most impact the spectra quality and the developed model's performance. To a lesser extent, the change of the optical length in the measuring instrument can also affect the reliability of the studied property estimation. Nevertheless, a suitable preprocessing scheme can partially correct this last parameter. As evidenced in the discussion of this chapter, there is no single solution to the issue of model robustness. Even in chapter 4, it was demonstrated that partial robustness of the model is achieved through data fusion modelling. Thus, as emphasized throughout the chapters of the manuscript, selecting the chemometric method that best meets the research need depends largely on the researcher's objective and the available information. This chapter addresses the model robustness issue regarding the external parameters associated with the NIR spectra acquisition.

The correction of the external parameters is usually accomplished independently, i.e., correcting the influence of one parameter at a time. When the robustness constraint is linked to instrumental changes, the widely used option to compensate for the impact of this parameter is the *a priori* correction, being the PDS method the most employed. However, this method has the disadvantage of requiring the transformation of the new spectra using a transfer function, being necessary to have enough reference samples (typically 20) to achieve an optimal function fit. In addition, the developed transfer function is usually not generalized and can be affected by other external parameters, including slight instrumental changes that were supposedly already included in it. Even when a generalized transfer function is developed for correcting different external parameters simultaneously, it fails to fully compensate for the impact of parameters having a non-linear incidence on the spectra quality, such as sample temperature and dynamic acquisition. As a result, a recalibration of the function is required whenever a significant deviation in model performance is evident due to an external parameter.

Another strategy for correcting the impact of several external parameters is calibrating a regression model for each G parameter. However, despite the promising results, its development and application are often impractical. Firstly, it requires a database for each parameter evaluated, rendering it an economically unfeasible alternative. Secondly, this strategy increases the complexity of analyzing, defining, and applying several regression models to predict a single property. Additionally, this strategy does not correct the impact of factors present in dynamic acquisition.

If reliable property estimation is intended for online monitoring applications, the strategies described in the previous paragraphs are limited. There is a high probability that the impact caused by several external

parameters occurs simultaneously in real-time property estimation applications. The generation of a global regression model using orthogonalization as a preprocessing method (OSC) partially corrects the effect that dynamic acquisition conditions have on the model accuracy. However, its robustness is not suitable enough. Based on the favorable outcome of the preprocessing method used in developing the generalized model, the combined use of EPO and DOP's orthogonalization methods was satisfactorily applied.

When employing the EPO method, the simultaneous correction of several external parameters is achieved from a single, generalized PLS model. The effectiveness of this method lies in utilizing a detrimental matrix D that describes the impact that the different parameters have on the studied property. The EPO method has the advantage that the impact description of several external parameters can be included in this matrix D simultaneously. Therefore, it could be inferred that the impact caused by dynamic acquisition is corrected when using a PLS_EPO model calibrated using a D-matrix that includes the impact of instrumental changes and sample temperature. Nonetheless, the factors associated with the dynamic acquisition conditions are not always easily identified and analyzed. For instance, the impact of the flow rate that leads to instantaneous sample changes during the spectra acquisition. Therefore, extending the model robustness to consider these factors is necessary. However, since it is difficult to measure the impact of these factors, this task is highly complex. The DOP method offers an effective solution to overcome this drawback.

This thesis implemented an integrated use of the EPO and DOP methods (see paper 5 appendix 5). As a result, a robust model that effectively corrects for the impact of all the external parameters studied was obtained. The integration of these two methods presents an advantage over the other correction methods by facilitating the incorporation in the D-matrix of any other external parameter impact. Consequently, the model robustness increases continuously by including the impact of external parameters evidenced during process monitoring and operation, such as feedstock and operating conditions changes. In addition, the integration of these two methods offers a relative simplicity in the model maintenance since it does not require a large volume of data or reference samples. For instance, the results shown in this chapter were achieved by updating the EPO model using a single NIR spectrum acquired under dynamic conditions. One limitation encountered when applying this strategy is the need to know the analyzed property's measured value to use the DOP method.

In summary, the integrated application of the orthogonalization methods showed that regardless of the acquisition conditions, stable or variable, steady or dynamic, and even with different types of instruments, the middle distillate properties prediction is reliable over the whole range of estimation evaluated. In addition to the model robustness achieved, a great advantage of using orthogonalization methods is that no further processing or transformation of the new spectra is required, facilitating the maintenance of the models over time.

# CHAPTER VI
# Model Deployment

# Chapter VI. Model Deployment

The research questions formulated in this thesis were addressed considering different evaluation scenarios. As a result, a robust and reliable estimation of middle distillates properties was achieved. This milestone accomplished through the thesis development raises the perspective of implementing the obtained models in daily operation analysis, including real-time monitoring. Therefore, it was decided to validate the results and conclusions drawn in the previous chapters by implementing the models in scenarios not considered during their calibration and testing.

This chapter is divided into three main sections. The first section presents the results obtained by applying each developed model to samples analyzed at steady-state and offline conditions. The second section shows the application of one of the developed models for monitoring a test conducted in a hydrocracking pilot plant. Finally, the third section summarizes the conclusions drawn from the analysis of the results.

## 1. Case 1: Offline and steady acquisition conditions

For this first case, NIR spectra were acquired on 26 total effluent samples obtained during a test conducted in 2021 in one of HCK's pilot plants located at the IFPEN in Solaize, France. Following the same acquisition protocol for steady-state conditions as described in the materials and methods section (see chapter 2), the spectra were acquired at a sample temperature of 60°C, using a Falcata Lab 6 immersion probe (reflectance) with a fixed optical length of 2 mm.

The samples were classified into four groups:

- Group 1: samples obtained by processing feedstocks under operating conditions and catalytic systems employed before when producing the samples used in the model calibration (7 samples);
- Group 2: samples obtained by employing a catalytic system not previously included in the database, but processing feedstocks already used (9 samples);
- Group 3: samples obtained by processing feedstocks not previously evaluated but using catalytic systems included in the database for generating the models (5 samples);
- Group 4: samples obtained by processing feedstocks and catalytic systems not included in the database (5 samples).

Eight different models were applied to the acquired spectra (see Table 24). For this first case, this chapter only shows the results and respective analysis of two diesel properties, namely the cetane number and the CFPP. The results of the remaining properties studied are reported in Appendix 7.

*Table 24. Description of the models applied to the 26 new total effluent samples*

| Model | Description | Acronym |
|---|---|---|
| 1 | NIR model | Mod_1 |
| 2 | NIR model with EPO_DOP correcttion | Mod_2 |
| 3 | NMR model | Mod_3 |
| 4 | NIR + NMR data fusion model | Mod_4 |
| 5 | NIR + PV_TE* data fusion model | Mod_5 |
| 6 | NIR + PV_NTE** data fusion model | Mod_6 |
| 7 | NIR + NMR + PV_TE data fusion model | Mod_7 |
| 8 | NIR + NMR + PV_NTE data fusion model | Mod_8 |

**PV_TE = Process Variables including Total Effluente properties*
**PV_NTE = Process Variables not including Total Effluente properties*

## 1.1 Diesel Cetane Number

Figure 17 compares the eight models' performance in estimating the diesel cetane number. The comparison analysis was performed regarding the SEP, the bias, the RMSEP and the $r^2P$. This figure corroborates the conclusions proposed in the previous chapters, validating the responses given to each research question. A general analysis reveals that the NIR base model (Mod_1) has a lower RMSEP than the reproducibility of the reference method (±3.6). Next, it can be corroborated that when using complementary information to the NIR spectra (NMR spectra and process variables), the model's performance is improved (lower RMSEP and higher $r^2P$). Finally, the robust model using only the NIR spectrum of the total effluent samples (Mod_2) has a better performance than the base NIR model.

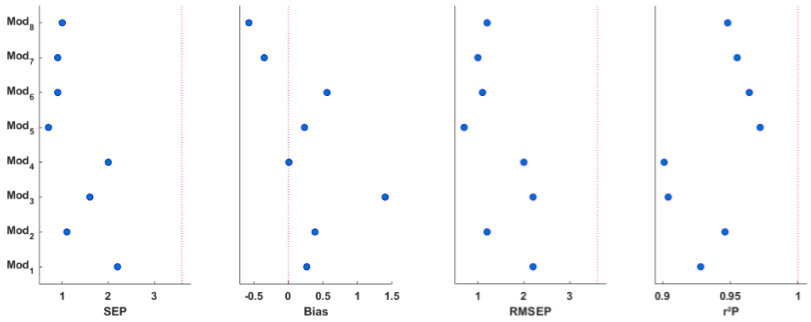



*Figure 17. Model performance comparison for predicting diesel Cetane Number in 26 new samples*

Figure 18 provides further details for the model performance analysis. This figure shows the parity plots corresponding to each model.

The NIR model successfully predicts 24 out of the 26 new samples within the reproducibility limits, representing a prediction effectiveness of 92%. From the two samples predicted out of limits, it can be observed that the one having the largest deviation corresponds to a sample obtained when processing a new feedstock. Conversely, all samples obtained using a new catalytic system were predicted within the reproducibility limits. When comparing the samples according to the defined group in the introduction of this

first case, those obtained from a new feedstock (blue squares) show a slightly larger deviation than those obtained using a new catalytic system (red diamonds). These results could be interpreted as if the chemical information extracted and exploited from the NIR spectra is better able to capture the changes in the catalytic system rather than in the properties of the feedstock. Nonetheless, the global performance of the NIR model is satisfactory in this study case.

An interesting trend observed in the estimations obtained with the NIR model is a positive bias for samples with a cetane number higher than 60, while a negative bias is observed for samples with a cetane number lower than 60. When analyzing this behavior regarding the different feedstock, total effluent properties and the operating conditions, it was found that this trend is strongly associated with the process conversion. This parameter is directly influenced by the process operating temperature. The higher the temperature, the higher the conversion. Diesel, being a heavier product than total effluent and kerosene, its molecular interaction is more impacted by temperature. Hence, the trend of the NIR model estimate can be explained. When the robust EPO model (Mod_2) is applied, this effect is corrected, and the cetane number prediction is improved. This result validates the importance and reliability of the model's robustness.

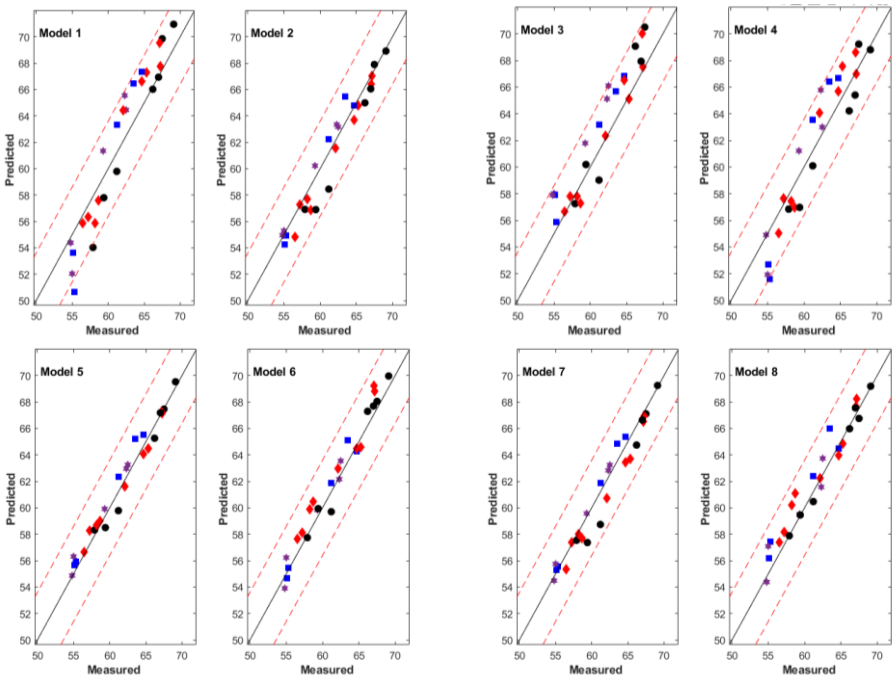



*Figure 18. Parity plot for model performance comparison in predicting diesel Cetane Number in 26 new samples. Red dotted lines: reproducibility limits of the reference method (±3.6). Legend: Black circles → Samples group 1. Red diamonds → Samples group 2. Blue squares → Samples group 3. Purple stars → Samples group 4*

The most accurate estimates were obtained with the data fusion models between spectroscopic information and process variables. The lowest RMSEP and the highest $r^2$ are obtained when the total effluent properties are used in the cetane number prediction (Mod_5). However, if cetane number estimation is desired for online process monitoring, this model cannot be applied due to the unavailability of real-time total effluent properties information. Instead, the data fusion model using the NIR spectra, the feedstock properties, and the operating conditions (Mod_6) could be used with comparable performance.

## 1.2 Diesel CFPP

As detailed in Chapters 3 and 4, the most complex diesel properties for modelling were the cold flow properties. This section shows the results of implementing the developed models for estimating the diesel CFPP of the 26 samples analyzed in this chapter. As a brief reminder, the results presented in the previous chapters indicated that while the base NIR model has an RMSEP close to the reference method reproducibility, the prediction bias and the $r^2P$ are not optimal, and some samples were predicted outside the reproducibility limits. However, when using data fusion, the prediction of diesel CFPP was improved.

Figure 19 shows the comparison between the 8 models evaluated. The same trend described in the previous paragraph can be observed. Once again, it can be validated that the synergistic use of supplementary information contributes to a more accurate description of this diesel property. The data fusion model involving spectroscopic information and process variables, including the total effluent properties, is the best performing model. Compared to the NIR model, this data fusion model significantly improved the $r^2P$ and the RMSEP (67% reduction). Unfortunately, similar to cetane number estimation, this data fusion model is limited for online monitoring applications due to the unavailability in real-time of the NMR spectra and the total effluent information. Nevertheless, the data fusion model that uses the NIR spectrum and process variables without considering the total effluent properties provides a satisfactory estimate of the diesel CFPP.

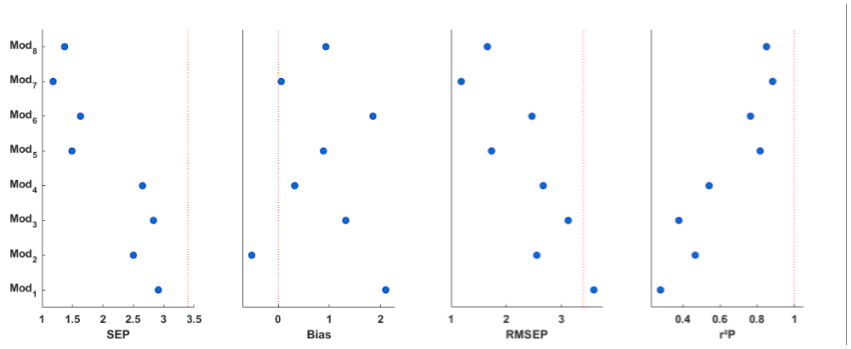



*Figure 19. Model performance comparison for predicting diesel CFPP in 26 new samples*

The parity plots shown in Figure 20 provide further details for a deeper analysis of the models' performance. From these plots, it can be emphasized that, except for the data fusion model described previously (Mod_7), all models fail to predict the sample with the highest CFPP value within the reproducibility limits. The value of the studied property corresponding to this sample (+4) exceeds the upper limit of model applicability (+2), explaining the poor prediction observed.

When analyzing the models' performance closely, it can be observed that out of the 26 samples evaluated, the NIR model predicts 20 of them within the reproducibility limits (77% prediction effectiveness). Five of the six samples estimated outside the limits correspond to samples acquired while processing feedstocks and evaluating catalytic systems not included in the initial database used for model calibration. Unlike the cetane number, the chemical information contained in the NIR spectra does not fully describe the impact of these

operational changes in the studied property. Regarding the robust EPO model, the improvement in property estimation is once again achieved. Compared to the NIR model, the prediction effectiveness percentage is improved (85%), the bias and RMSEP are reduced by 75% and 29%, respectively, and higher homoscedasticity of the model is observed.

When applying the data fusion model employing NIR spectra and process variables, without including the total effluent properties (Mod_6), 25 out of the 26 samples analyzed are predicted within the reproducibility limits. This better property estimation is reflected in the RMSEP reduction and the $r^2P$ improvement. Such promising results suggest the possibility of deploying this model with relative reliability in online monitoring of the diesel CFPP behavior. Nevertheless, it is recommended to analyze the possible causes of the positive bias observed in this model. It is noteworthy that this bias is slightly corrected when two parameters related to the total effluent are included, i.e., conversion and the simulated distillation temperature used to recover the 30% of the diesel T30 (Mod_5). This bias reduction could be associated with the impact that the quality of the feedstock and the catalytic system have on the process conversion. Therefore, this parameter could provide complementary information to the model, helping to explain better the impact of the new operating conditions on the diesel CFPP.

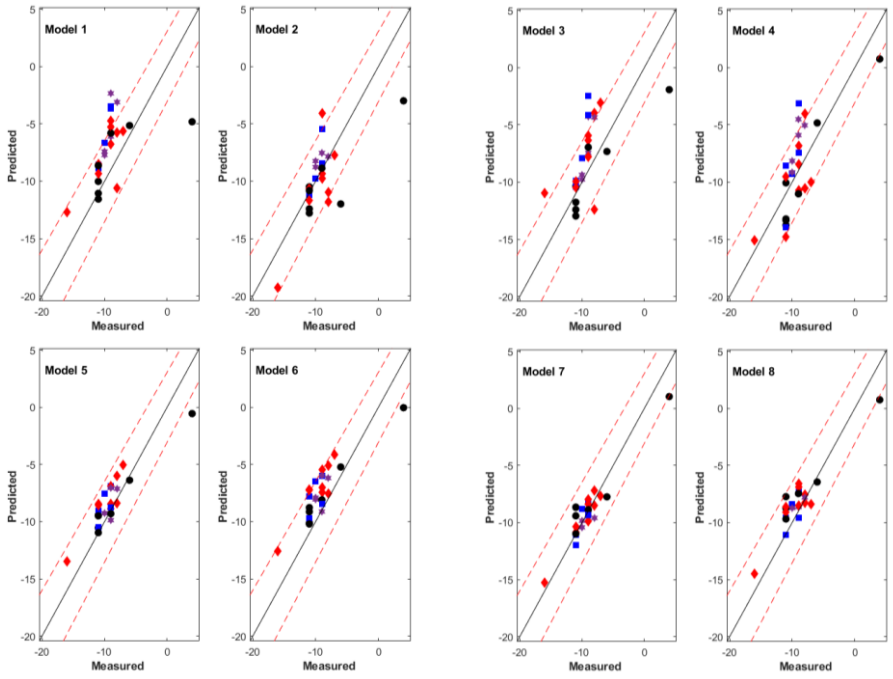



*Figure 20. Parity plot for model performance comparison in predicting diesel CFPP in 26 new samples. Red dotted lines: reproducibility limits of the reference method (±3-0.06*CFPP). Legend: Black circles → Samples group 1. Red diamonds → Samples group 2. Blue squares → Samples group 3. Purple stars → Samples group 4*

Finally, as previously discussed, the best-performing model was the data fusion model using the samples' NMR spectra and the process variables as complementary information. The data description provided by the NMR contributes to describing the behavior of the analyzed property more accurately. When using this model, the prediction effectiveness is 100%. Compared to the NIR model, this model significantly reduces the SEP (59%), the bias (79%), and the RMSEP (67%).

The results discussed in this first case are relevant when validating the conclusions presented in the previous chapters. It is worth stressing that the samples used for model calibration were collected from tests performed in the IFPEN pilot plants between 2013 and 2018 using different feedstocks and operating conditions. In addition, the spectra were acquired on these samples during the second semester of 2020. The 26 new samples evaluated in this section were obtained in 2021 using feedstocks and operating conditions different from those used in the previous years. Moreover, the spectra acquisition on these samples was conducted nearly a year after (2021) acquiring the initial spectra. Despite the variability described concerning the process and dates of sample analysis, the performance of the models was satisfactory.

## 2. Case 2: Pilot plant test monitoring

The results discussed throughout this manuscript have validated the feasibility of estimating middle distillate properties independently of the total effluent distillation. This alternative product characterization minimizes the response times. With a more opportune response, process analysis and decision-making are optimized. Hence, one of the further objectives of the thesis is to achieve the digital monitoring of the HCK process. This section shows the implementation of the developed work in the online monitoring of a test conducted in one of the IFPEN's pilot plants. A general description of the experimental setup in the pilot plants is given below to help the reader to have a more precise context of this study case. Subsequently, the conditions for acquiring the spectra and the description of the models used in this case are summarized. Finally, the results obtained are shown and discussed.

In broad terms, a test in the HCK process pilot plants starts with the *in-situ* catalyst sulfiding using a gas oil spiked with aniline and dimethyl disulfide (DMDS). Following this step, the operating conditions are set, and the feedstock to be processed is injected. From this point on, the stability of the pilot unit is continually supervised. Monitoring the unit stability is crucial to ensure the representability of the total effluent samples collected. The unit stabilization is tracked by monitoring the total effluent density. This property is measured offline every 24 hours. The pilot unit operation is considered stable when the density variation is lower than 0.0005 g/cm3. It should be noted that during the test in the pilot plant, different operational changes related to the feedstock, pressure, temperature, and residence time are evaluated. Once the stability is reached, the total effluent is collected to be further distilled for obtaining and characterizing the cuts. The results obtained from the characterization are used to evaluate the process performance.

This second study case was conducted to evaluate the feasibility of online monitoring of the HCK process by implementing real-time NIR spectra acquisition on a pilot plant. To this end, a transmittance flowcell with an optical path of 1 mm was installed at the process reactor outlet. The spectrometer (Metrohm) used for the spectra acquisition is detailed in chapter 2. The acquisition frequency of the NIR spectra on the total effluent was 5 minutes.

The monitoring of the pilot plant test focused on two property estimations. First, the total effluent density was estimated to evaluate the pilot unit stability. This property was chosen for the daily measurements of the density on the total effluent giving access to an interesting amount of reference values to validate the NIR approach. A model was calibrated from a data set of 170 samples following the robust Modelling principles described in Chapter 5 to estimate this property. This model had an RMSEC and RMSECV of 0.0018 g/mL and 0.0020 g/mL, respectively. The model's $r^2C$ (0.994) and $r^2CV$ (0.991) were optimal. The model's performance was validated using an external data set of 65 samples, resulting in an RMSEP of 0.0019 and an $r^2P$ of 0.991. The described model corresponds to a PLS model of 9 LV and 7 EPO components. Simultaneously to the total effluent density estimation, the diesel cetane number was predicted using the PLS_EPO model described in chapter 5. This property's prediction was performed to analyze its behavior throughout the test regarding the operational changes evaluated. Figure 21 shows the results obtained from the monitoring conducted for approximately 44 days.

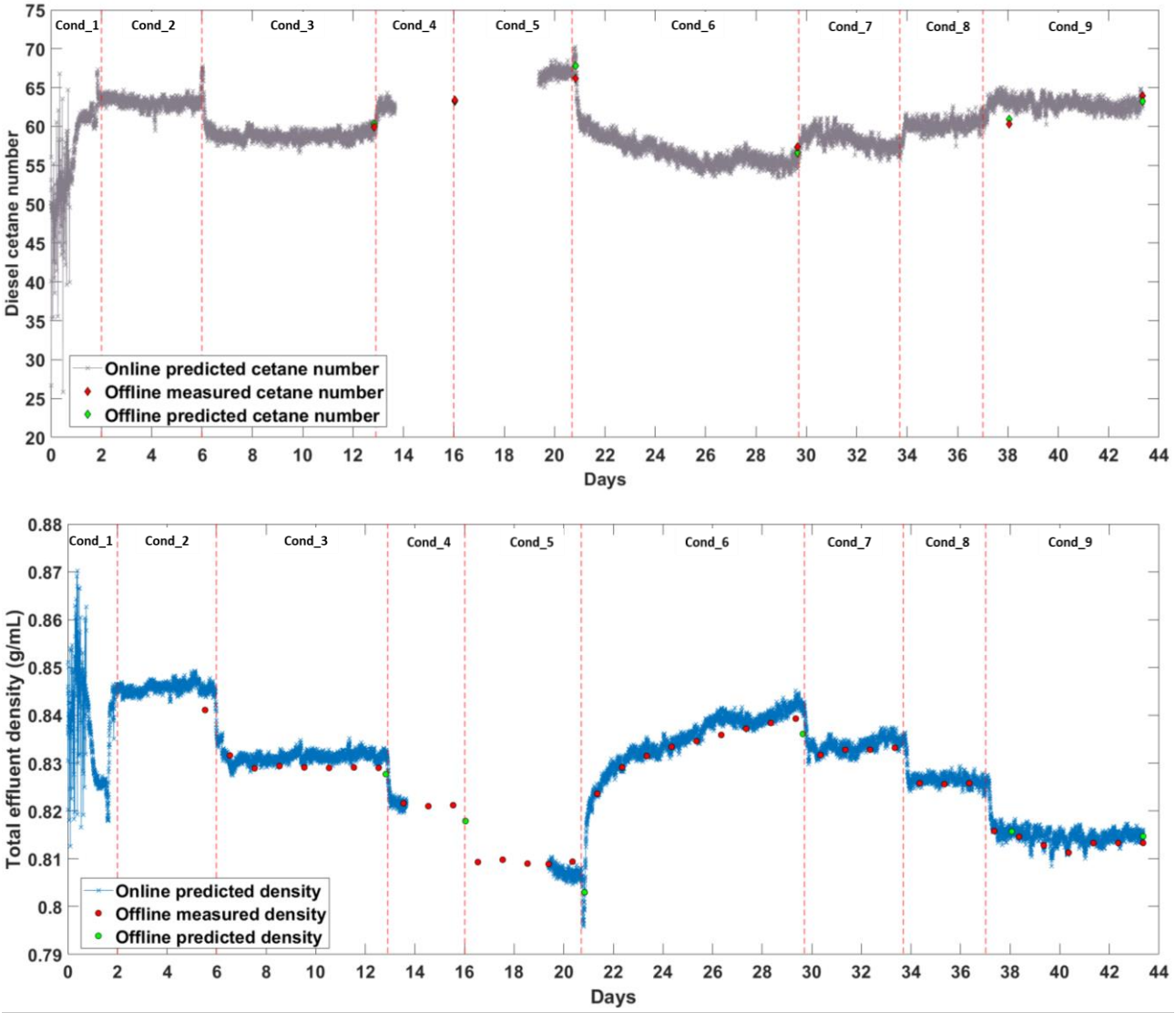



*Figure 21. Online HCK process monitoring. Legend → Cond_1: catalyst sulfiding, Cond_2: Feedstock A injection, Cond_3: Feedstock B injection, Cond_4: Operating temperature increase, Cond_5: Operating temperature increase, Cond_6: Feedstock C injection, Cond_7: Operating temperature increase, Cond_5: Operating temperature increase, Cond_9: Operating temperature increase*

Before presenting the analysis of the results, it is necessary to mention that the lack of properties prediction between days 14 and 19 was due to an unintentional shut-down of the spectrometer.

## 2.1 Online total effluent density estimation

The online density predictions (blue line Figure 21) were compared with the measured value of this property every 24 hours (red circles Figure 21). The first remark that can be made is the satisfactory prediction of the total effluent density over the entire test. Compared to the offline density measurement conducted every 24 hours, it can be seen in Figure 21 that the online density estimation is accurate and correctly predicts the property behavior. In addition, the online density monitoring enabled timely observation of the impact on process stability caused by catalyst sulfiding (days 0 - 2), injection (day 2) and feedstock changes (days 6 and 21), as well as changes in operating conditions (days 13, 16, 30, 34, and 37).

Besides the accurate prediction of the process stability trend, the estimated and measured value difference is close to the reference reproducibility for this property in products such as total effluent (0.0011 g/cm3). When disregarding the offline density measurements made on days 6 and 27, the maximum difference found was 0.0024 g/cm3, while the minimum difference was 0.0001 g/cm3, with an average difference of 0.0014 g/cm3. The estimated density on day 6 showed a difference of 0.0036 g/cm3 compared to the measured value. The accumulation of the total effluent sample under unstable conditions during the first two days could explain this difference. On the other hand, the prediction for day 27 was the one that presented the greatest difference (0.0042 g/cm3) without any particular explanation for this discrepancy. Despite these occasional differences, implementing the density model for monitoring the process was successful.

Online process monitoring offers the possibility of reducing the time currently used to establish the unit stability. Moreover, an opportune analysis of the impact of different parameters on the process stability and product quality can be achieved. The response time optimization could be reflected in more efficient and expeditious process research.

## 2.2 Online diesel cetane number estimation

Unlike the total effluent density, the measured value of the diesel cetane number is conditioned to the total effluent distillation and the respective laboratory analysis. Six total effluent samples corresponding to the operational changes evaluated were produced during this test. These samples were subsequently distilled to recover the diesel cut, on which the cetane number was measured. It should be noted that the time gap between collecting each total effluent sample and measuring the diesel cetane number was between 3 and 4 weeks. Keeping this in mind, the diesel cetane number values predicted online (black line Figure 21) could only be validated with the measurement of these six diesel cuts analyzed (red diamond Figure 21).

Similar to the total effluent density, the estimate of the diesel cetane number consistently reflects the

behavior of this property as a function of the operational parameters evaluated during the test. For example, the higher the operating temperature at the same feedstock, the higher the diesel cetane number. Compared to the six measured values, the estimate of the diesel cetane number is accurate, with a minimum and maximum deviation of 0.2 and 1.6, respectively. These online estimates are not only reliable but also rapid. While it took 3 to 4 weeks to know the measured value of the cetane number after collecting the total effluent samples, the estimated value using the developed robust model enables the researcher and the process operators to know this value in a couple of minutes.

In addition to the online process monitoring, a complementary analysis was performed using the six total effluent samples accumulated during the test. The analysis compared the robust model performance when predicting the total effluent density and the diesel cetane number under offline and online conditions. For the offline prediction, the NIR spectra were acquired under stationary conditions, at a constant sample temperature of 60°C and a fixed optical length of 2 mm using a reflectance instrument (see chapter 3). For the online prediction, the NIR spectra acquired during the process monitoring (possibility of variability related to the sample temperature and flowrate) with a transmittance instrument of 1 mm optical length were utilized. Figure 21 shows the capacity of the robust models to reliably predict these two properties at different acquisition conditions (blue line vs. green circles for density, black line vs. green diamond for diesel cetane number).

# 3. Concluding remarks

The results shown in this chapter validated the main findings and conclusions drawn in the previous chapters when using a test data set with information not included in the properties modelling. It was confirmed that predicting the middle distillates properties from the chemical information contained in the total effluent NIR spectra is feasible. It was also corroborated that the properties estimation can be improved when complementary information to the NIR spectrum, such as NMR and process variables, is used synergistically. Lastly, it was validated that developing a robust model is fundamental to ensuring reliable property estimation under different evaluation conditions over time.

Obtaining robust models using orthogonalization methods was a crucial milestone in developing the thesis. The results showed that this model corrects the impact of temperature at the macro and micro levels. To give an example, let us take the diesel cetane number estimation. At the macro level, the model corrects the impact of sample temperature on the spectra quality, providing a reliable property estimate. Regarding the micro level, the impact that the operating temperature has on the property estimation can be outlined. When the reference NIR model is used to estimate the diesel cetane number, a positive and negative bias is observed, explained by this external parameter influence. On the contrary, using the robust model, the

impact is corrected by eliminating this parameter's biases on the property estimation.

Compared to the robust NIR model, the data fusion models using process variables show a more accurate prediction when changes in process operation occur. This better performance could be interpreted as a greater capacity to correct the impact of operating conditions on the property estimation. However, this comparison is valid as long as the NIR spectrum acquisition conditions are the same as those used to obtain the spectra used in the model calibration. Recalling the results shown in chapter 5, it is evident that the external parameters related to the spectra acquisition significantly affect the models' performance. If the data fusion models were deployed under different NIR spectrum acquisition conditions, the information provided by this data block could be biased, affecting the model performance. In that case, the robust model developed from orthogonalization methods would give a more accurate estimate. This simplified analysis raises the perspective of finding a methodology that allows the complementary use of orthogonalization and data fusion methods. The synergistic use of the advantages that each approach brings to the table could result in a more robust model.

The robust modelling done in this thesis was applied to a case study of online monitoring of the HCK process. Compared to the measured values of the property used to determine the process stability, the prediction model showed satisfactory performance. The robust model developed provided a reliable estimate over the entire test monitoring, thus demonstrating its high consistency. The results obtained were highly promising. The model consistently and reliably estimated the total effluent density, even when changes in process variables such as feedstock quality and reaction temperature occurred. Therefore, the online process monitoring using the robust model enabled the opportune analysis of different operating conditions' impact on the process stability, helping to have a more accurate and timely process analysis.

One of the major advantages of the alternative investigated in this thesis is the property estimation of different products from a single NIR spectrum. In addition to the total effluent density prediction for the process stability analysis, the middle distillate properties could be reliably estimated, particularly the diesel cetane number. It should be pointed out that the correct property prediction using the robust model is achieved either from spectra acquired under dynamic or steady-state conditions. These results reinforce the power and versatility of orthogonalization methods in the robust model generation.

Although the robust models developed had remarkable performance, it should be noted that these models are not necessarily flawless, and their performance may deteriorate in applications where the impact of certain external parameters has not been included in the model calibration. Therefore, it is important to maintain these models periodically. This thesis showed that the integrated use of the EPO and DOP methods gives a practical manner for model updating.

Finally, it is worth highlighting that the response times in the middle distillate characterization could be reduced from several weeks to a few minutes. For example, the measured values of the diesel cetane number used to validate the robust model performance were available after several weeks: time employed to

perform the total effluent distillation to recover the diesel cut and its respective characterization. On the contrary, the estimation of the diesel cetane number could be achieved in a couple of minutes using robust models. The results obtained raise the prospect of integrating the models into the process control system.

# General Conclusions and Perspectives

# Conclusions

The hydrocracking process is an ongoing research subject. Characterizing the products obtained from this process, particularly middle distillates, is essential in the research work. However, the analytical workflow traditionally employed is both time- and volume-consuming. While the standard laboratory analyses contribute to sub-optimal response times, the total effluent distillation gives the major time constraint to obtain the physical cuts. This property was chosen for the daily measurements of the density on the total effluent giving access to an interesting amount of reference values to validate the NIR approach. Hence, this thesis addressed the need previously outlined. Considering that the major constraint is the total effluent distillation, the main challenge was developing an alternative to characterize the middle distillates reliably and robustly, avoiding the distillation step. That is, to achieve the product characterization without relying on the physical sample. In addition, this alternative had to demonstrate the potential to be applied in real-time process monitoring.

The combined use of analytical techniques and chemometric methods addressed the described need. Being an analytical technique that overcomes the limitations of response time and sample volume required, NIR spectroscopy was the focus of the thesis research. The first research problem was to evaluate the feasibility of using the total effluent spectroscopic information to estimate middle distillate properties. The results obtained validated this alternative in characterizing the HCK process products.

The middle distillates characterization was achieved by calibrating PLS models from the NIR spectra acquired on the HCK total effluent. The prediction errors were close to the reproducibility limits of the reference methods regularly used. The best-estimated property was the cetane number of diesel and kerosene. On the contrary, the diesel cold flow properties were the most challenging to estimate reliably. Although the overall prediction error of these properties was close (equal to or slightly higher) than the reproducibility of the reference methods, the homoscedasticity of the developed model was not optimal. This limited performance was reflected in the $r^2P$, the bias, and the prediction of some samples outside the reproducibility limits. A more detailed analysis of the results revealed that this model performance was mainly due to two factors. First, the complexity of the diesel cold flow properties to be described from only the NIR spectra of the total effluent, and second, the influence that operating variables could have on them.

Data fusion modelling solved the performance limitations of the NIR regression models, especially those for estimating the diesel cold flow properties. Three data blocks were used in this modelling approach: the (i) NIR and (ii) NMR spectra of the total effluent and (iii) the process variables. The appropriate data fusion approach improved the homoscedasticity, squared correlation coefficient, bias, prediction error, and the number of predicted samples within reproducibility limits. While it was validated that combining the relevant

information provided by each data block contributes to the models' performance improvement, the selection and use of the most descriptive variables provide further enhancement. By employing the appropriate variable selection methods in each block, optimal performance of the models was evidenced, achieving the cold flow properties estimation of all samples within the reproducibility limits. In addition, the variable selection has improved the properties estimation and facilitated the identification, analysis, and validation of the parameters that impact the studied properties. Consequently, a comprehensive analysis of the process could be accomplished.

The developed models showed satisfactory performance in estimating each studied property. A case study was implemented to validate the models' performance when considering total effluent samples not included in the initial calibration and validation datasets. These samples were obtained from processing feedstocks and catalytic systems not previously evaluated in the models. The samples and their respective spectroscopic information were acquired later than the models' development. Despite the differences described, the results confirmed the aforementioned findings, i.e., the prediction of middle distillates from total effluent spectroscopic information is feasible, and the estimation of properties is improved by using data fusion and variable selection.

Despite the results obtained in the final models' validation, their performance can be affected by external parameters associated with the spectral acquisition conditions. Therefore, the development of robust models was a fundamental part of the thesis research. Orthogonalization methods, particularly EPO and DOP, were used in the robust modelling. The models obtained provided reliable estimates of middle distillate properties under different acquisition conditions. The external parameters corrected for were sample temperature, changes in the spectral acquisition instrument, and external factors involved in the dynamic acquisition of the spectra.

The robust models were developed to ensure reliable property estimation under different evaluation conditions, especially in the dynamic spectra acquisition. Furthermore, the model robustness enables its deployment in online monitoring applications. Thus, a second case study was implemented to validate the performance of the developed robust models in the HCK process monitoring. The results showed the robustness of the developed total effluent density model. Over the entire test monitoring, the satisfactory model performance was consistent, providing insights into the influence that different operational changes had on the process stability. Furthermore, the estimated value of the density presented deviations close to the reproducibility of the reference method, ensuring the prediction accuracy. Another important fact to highlight is the capability of the model to predict the density of the total effluent under steady-state and dynamic conditions with satisfactory reproducibility. The diesel cetane number prediction validated the importance of model robustness. This property was also estimated during the online process monitoring with

similar results to those discussed already.

The models using NMR spectra and total effluent properties are limited for online process monitoring due to the unavailability of this information in real-time. In contrast, with satisfactory performance, the robust NIR model and the data fusion models between the NIR spectra, feedstock properties and operating conditions can be implemented in real-time estimations.

The research conducted in this thesis offers a reliable and robust alternative for optimizing the product characterization involved in the research and development of the HCK process. This alternative is a further step in research about generating and applying strategies for optimizing product characterization. Though the response time optimization of this task via spectroscopic techniques is extensively known, this option is generally conducted using the information acquired on the sample being analyzed. This characterization strategy is still dependent on the time employed to obtain the physical samples to be analyzed, which in this thesis is the total effluent distillation for obtaining the middle distillates. Additionally, if it is necessary to characterize more than one product, the strategy outlined above would require the spectra of each analyzed product. Implementing the alternative investigated in this thesis achieves further time and cost optimization. When using the spectroscopic information of the total effluent to estimate the middle distillate properties, the time constraint given by the distillation is overcome. Furthermore, this alternative has the advantage of using a single spectrum to estimate the properties of several products.

As previously discussed, compared to standard reference methods for product characterization, one advantage of combining chemometrics with analytical techniques is the improved response time. Another advantage is the possibility of extending the property estimation range. For example, for diesel cetane number, the applicability range of the standard norm is between 40 and 56, whereas with the developed chemometric models, this property can be reliably estimated in a range between 36 and 70. This latter advantage is conditioned by the database utilized in developing the models.

Using a representative database in model calibration is important to have an accurate and reliable description over the entire range of model evaluation. For instance, for some properties, particularly the diesel cold diesel flow properties and kerosene flash point, it was observed that the dataset was more populated at the upper end of the model application range. This non-homogeneous data population makes it difficult to explain the property reliably since samples from these sparsely populated areas far from the data set center could behave as outliers, even if they are not. Therefore, it is advisable to use a complete and homogeneous populated database to obtain better-performing base models.

The analysis and detection of anomalous data are also important during model calibration to ensure that the models properly describe the behavior of the property being studied. In addition, this analysis should be

performed during model deployment using new data to determine if a poor model estimate is due to a model constraint or if it is indeed due to the outlier nature of the data. The most common method used in this thesis was the combined analysis of the Qresidual and Hotelling $T^2$ tests.

The results obtained through the thesis development led to conclude that the investigated alternative can be reliably applied in real-time applications, resulting in optimized process analysis, either by its online monitoring or by estimating the middle distillate properties in real-time.

Online process stability monitoring ensures that the decision-making is timely. As a result, the researcher can optimize the plant's operating time. For instance, the pilot plant stability is usually established by analyzing the total effluent density variability. This value is measured every 24 hours. To reliably determine stability using these measured values, at least three density points are required, implying that the decision-making time could be around two or three days. Through online monitoring, the pilot unit stability could be determined earlier, reducing operating times. Another advantage resulting from online monitoring is the early identification of unplanned deviations, enabling timely and accurate process adjustments.

Estimating middle distillate properties in real-time has two main advantages. The first is the optimized process analysis that the researcher can perform by readily knowing the effect of operating conditions on product properties. The average response time for knowing this information following the traditional workflow can be several weeks. With the work developed in this thesis, this time can be reduced to a few minutes. The second advantage is related to cost optimization. Having real-time knowledge of the process operation behavior and its impact on product quality, the researcher can decide on the actual analytics needed for the research.

When implementing the models in real-time monitoring and analysis applications, it is important to be aware of operating schemes whose products tend to deposit solid residues at low temperature on the instrument used for spectrum acquisition, affecting the light transmittance. This phenomenon mostly occurs when obtaining high paraffinic content products during the winter. Therefore, it is recommended to ensure a minimum safe temperature in the instrumentation related to spectrum acquisition, either through proper insulation or a complementary heating system.

Finally, it is important to emphasize the importance of periodic model maintenance to ensure optimal performance over time. Different alternatives can be used to perform this task. However, for online monitoring and real-time property estimation applications where the dynamics of the operation can rapidly deteriorate the performance of the models, the integration of the EPO and DOP methods becomes an effective and easy-to-implement alternative.

# Perspectives

The research findings validated the feasibility of optimizing the research and development of the HCK process. However, this work still has a long way to go. In consequence, from the conclusions drawn, it is recommended that the following perspectives be taken into account:

- **Integral model deployment in real-time process analysis and monitoring:** the results obtained in the process monitoring case study demonstrated the benefits of the investigated alternative. However, its implementation and definitive use are not yet complete. The remaining work is related to developing computational tools and interfaces for simplified use of the models. This work involves disclosing and teaching each person involved in the research labor the model handling and the interpretation of results.
- **Model prediction integration in the process control system:** the next step in optimizing the HCK process research is to employ the information obtained from the real-time estimations in the process control. As the thesis was developed, most of the principles used today in applying process analysis technology (PAT) and multivariate statistical process control (MSPC) were employed. Based on the research conducted, these methodologies could be implemented in the process operation.
- **Protocol development for model evaluation and maintenance:** The robust models developed efficiently estimate properties under different analysis conditions. However, model suitability may deteriorate over time due to operational changes or external parameters not being considered. For this reason, it is imperative that the models' performance is validated regularly and that an optimal methodology for models' maintenance is in place. For the case of real-time process monitoring and control, the orthogonalization method DOP has shown a good performance.
- **Development of regression models for estimating other relevant properties of middle distillates and other products obtained from the HCK process:** 7 middle distillates properties were studied in this thesis (4 for diesel and 3 for kerosene). However, there are other properties of these cuts that may be of interest in the HCK process research. Similarly, the properties of other products obtained, such as gasoline, unconverted oil (UCO), and dewaxed oil, may also be of interest. Thus, applying the methodology investigated in this thesis is recommended to analyze new streams or properties.
- **Extrapolate the work developed to other refining processes:** The work of this thesis focused on the HCK process. Nevertheless, the results obtained offer the possibility of using the methodology investigated in this thesis in other processes or industries when the same goal of optimizing the analytical workflow is sought.

# References


1. Bernard P. Tissot and Dietrich H. Welte, *Petroleum Formation and Ocurrence*. SPRINGER-VERLAG BERLIN AN, CHAM (1984).
2. B.P. Tissot and D.H. Welte, "Classification of Crude Oils", in: *Petroleum Formation and Occurrence,* Ed by B.P. Tissot and D.H. Welte, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 369–377 (1978).
3. Fuels Europe, *Statistical Report 2018*, Belgium (2018).
4. The Oxford Institute for Energy Studies and University of OXFORD, *The Light Sweet-Medium Sour Crude Imbalance and the Dynamics of Price Differentials* (2019).
5. A. Marafi, H. Albazzaz and M.S. Rana, "Hydroprocessing of heavy residual oil: Opportunities and challenges", *Catalysis Today*, **329** (2019).
6. M.S. Rana, "Heavy Oil Refining Processes and Petrochemicals: A Role of Catalysis", *Recent Adv Petrochem Sci*, **2** (2017).
7. R. Elshout, J. Bailey, L. Brown and P. Nick, "Upgrading the bottom of the barrel", *Hydrocarbon Processing* (March 2018).
8. Maureen Bricker, Vasant Thakkar, John Petri, "Hydrocracking in Petroleum Processing", in: *Handbook of Petroleum Processing 2014*, pp. 1–35.
9. Hydrocarbon Publishing Company, "WORLDWIDE REFINERY PROCESSING REVIEW - 2Q2017" (2017).
10. Frank (Xin X.) Zhu, Richard Hoehn, Vasant Thakkar, Edwin Yuh, "Description of Hydrocracking Process", in: *Hydroprocessing for Clean Energy,* Ed by F.X.X. Zhu, R. Hoehn, V. Thakkar and E. Yuh, John Wiley & Sons, Inc, Hoboken, New Jersey, pp. 51–78 (2017).
11. R.G. Tailleur, "Hydrocracking catalyst to produce high quality Diesel fraction", in: *Scientific bases for the preparation of heterogeneous catalysts. Proceedings of the 8th International Symposium, Louvain-la-Neuve, Belgium, September 9-12, 2002,* Ed by É.M. Gaigneaux, D.E. de Vos and P. Grange, Elsevier, Amsterdam, pp. 321–329 (2002).
12. S. Parkash, ed., *Refining processes handbook*. Gulf Professional Pub, Amsterdam, Boston (2003).
13. J.C. Vivas-Báez, A. Servia, G.D. Pirngruber, A.-C. Dubreuil and D.J. Pérez-Martínez, "Insights in the phenomena involved in deactivation of industrial hydrocracking catalysts through an accelerated deactivation protocol", *Fuel*, **303** (2021).
14. H. Toulhoat and P. Raybaud, eds., *Catalysis by transition metal sulphides. From molecular theory to industrial application*. Éd. Technip, Paris (2013).
15. E. Iplik, I. Aslanidou and K. Kyprianidis, "Hydrocracking: A Perspective towards Digitalization", *Sustainability*, **12**, 17 (2020).
16. J.M. Newsam, "High Throughput Experimentation (HTE) Directed to the Discovery, Characterization and Evaluation of Materials", *Oil & Gas Science and Technology - Rev. IFP*, **70**, 3 (2015).
17. ASTM D445-97, "Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity)". ASTM International, West Conshohocken, PA.
18. ISO 20846, "Petroleum products — Determination of sulfur content of automotive fuels — Ultraviolet fluorescence method". ISO International (2011).
19. ASTM D5291, "Standard Test Methods for Instrumental Determination of Carbon, Hydrogen, and Nitrogen in Petroleum Products and Lubricants". ASTM International, West Conshohocken, PA (2007).
20. ASTM D 3238 - 95, "Standard Test Method for Calculation of Carbon Distribution and Structural Group Analysis of Petroleum Oils by the n-d-M Method". ASTM International, West Conshohocken, PA (2000).
21. ASTM D1218 - 12, "Standard Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids", https://www.astm.org/Standards/D1218.htm.
22. ASTM S 1747 - 99, "Standard Test Method for Refractive Index of Viscous Materials". ASTM International, West Conshohocken, PA (2004).
23. ASTM D 7213-15, "Standard Test Method for Boiling Range Distribution of Petroleum Distillates in the Boiling Range from 100 °C to 615 °C by Gas Chromatography". ASTM International, West Conshohocken, PA (2015).
24. ASTM D2887 - 19ae1, "Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography", https://www.astm.org/Standards/D2887.htm.
25. ASTM D613-01, "Test Method for Cetane Number of Diesel Fuel Oil". ASTM International, West

Conshohocken, PA (2001).
26. D02 Committee, “Test Methods for Flash Point by Pensky-Martens Closed Cup Tester”. ASTM International, West Conshohocken, PA.
27. D02 Committee, “Test Method for Smoke Point of Kerosine and Aviation Turbine Fuel”. ASTM International, West Conshohocken, PA.
28. NF EN ISO 3015 (2019-05-29), “Petroleum and related products from natural or synthetic sources — Determination of cloud point”.
29. ASTM D5949, “Standard Test Method for Pour Point of Petroleum Products (Automatic Pressure Pulsing Method)”. ASTM International, West Conshohocken, PA.
30. 2016-03-24 (NF EN 116), “Diesel and domestic heating fuels - Determination of cold filter plugging point - Stepwise cooling bath method”.
31. F. Mabood, S.A. Gilani, M. Albroumi, S. Alameri, M.M. Al Nabhani and F. Jabeen, “Detection and estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with multivariate methods”, *Fuel*, **197** (2017).
32. M.G. Nespeca, R.R. Hatanaka, D.L. Flumignan and J.E. de. Oliveira, “Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis”, *Journal of Analytical Methods in Chemistry* (2018).
33. M.R. Monteiro, A.R.P. Ambrozin, M. da Silva Santos, E.F. Boffo, E.R. Pereira-Filho, L.M. Lião and A.G. Ferreira, “Evaluation of biodiesel-diesel blends quality using 1H NMR and chemometrics”, *Talanta*, **78**, 3 (2009).
34. R.M. Balabin, E.I. Lomakina and R. 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 (2011).
35. C.I. Rocabruno-Valdés, L.F. Ramírez-Verduzco and J.A. Hernández, “Artificial neural network models to predict density, dynamic viscosity, and cetane number of biodiesel”, *Fuel*, **147** (2015).
36. C.T. Pinheiro, R. Rendall, M.J. Quina, M.S. Reis and L.M. Gando-Ferreira, “Assessment and Prediction of Lubricant Oil Properties Using Infrared Spectroscopy and Advanced Predictive Analytics”, *Energy Fuels*, **31**, 1 (2017).
37. J.C. Lindon, G.E. Tranter and J.L. Holmes, eds., *Encyclopedia of spectroscopy and spectrometry*. Academic Press, San Diego, CA (2000).
38. M. Blanco and I. Villarroya, “NIR spectroscopy: A rapid-response analytical tool”, *TrAC Trends in Analytical Chemistry*, **21**, 4 (2002).
39. Herausgegeben von P. Williams. K. Norris, “Near-Infrared Technology in the Agriculture and Food Industries. Herausgegeben von P. Williams und K. Norris. 330 Seiten, zahlr. Abb. und Tab. American Association of Cereal Chemists, Inc., St. Paul, Minnesota, USA, 1987. Preis”, *Nahrung*, **32**, 8 (1988).
40. K.I Hildrum, T. Isaksson, T. Naes and A. Tandberg, *Near infra-red spectroscopy. Bridging the gap between data analysis and NIR applications*. Ellis Horwood, New York, London, Toronto etc. (1992).
41. E. Hatzakis, *Nuclear Magnetic Resonance (NMR) Spectroscopy in Food Science. A Comprehensive Review* (2018).
42. B. Caballero, P. Finglas and F. Toldrá, *Encyclopedia of Food and Health*, Oxford (2016).
43. M.K. Moro, F.D. dos Santos, G.S. Folli, W. Romão and P.R. Filgueiras, “A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy”, *Fuel*, **303** (2021).
44. A.F. Parisi, L. Nogueiras and H. Prieto, “On-line determination of fuel quality parameters using near-infrared spectrometry with fibre optics and multivariate calibration”, *Analytica Chimica Acta* (1990).
45. S. Garrigues, J.M. Andrade, M. de La Guardia and D. Prada, “Multivariate calibrations in Fourier transform infrared spectrometry for prediction of kerosene properties”, *Analytica Chimica Acta*, 317 (1995).
46. J.M. Andrade, S. Muniategui, P. Lopez-Mahia and D. Prada, “Use of multivariate techniques in quality control of kerosene production”, *Fuel*, **76**, 1 (1997).
47. M.I.S. Sastry, A. Chopra, A.S. Sarpal, S.K. Jain, S.P. Srivastava and A.K. Bhatnagar, “Determination of Physycochemical Properties and Carbon-Type analysis of Base Oils Using Mid-IR Spectroscopy and Partial Least-Squares Regression Analysis”, *Energy & Fuels*, 12 (1998).

48. N. Zanier-Szydlowski, A. Quignard, F. Baco, H. Biguerd, L. Carpot and F. Whal, “Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy”, *Oil & Gas Science and Technology - Rev. IFP*, **54**, 4 (1999).
49. H. Chung, M.-S. Ku and J.-S. Lee, “Comparison of near-infrared and mid-infrared spectroscopy for the determination of distillation property of kerosene”, *Vibrational Spectroscopy*, 20 (1999).
50. M. Kim, Y.-H. Lee and C. Han, “Real-time classification of petroleum products using near-infrared spectra”, *Computers and Chemical Engineering* (2000).
51. R. Meusinger and R. Moros, “Determination of octane numbers of gasoline compounds from their chemical structure by 13C NMR spectroscopy and neural networks”, *Fuel*, **80**, 5 (2001).
52. M. Gómez-Carracedo, J. Andrade, M. Calviño, E. Fernández, D. Prada and S. Muniategui, “Multivariate prediction of eight kerosene properties employing vapour-phase mid-infrared spectrometry☆”, *Fuel*, **82**, 10 (2003).
53. B. Basu, G.S. Kapur, A.S. Sarpal and R. Meusinger, “A Neural Network Approach to the Prediction of Cetane Number of Diesel Fuels Using Nuclear Magnetic Resonance (NMR) Spectroscopy”, *Energy Fuels*, **17**, 6 (2003).
54. A. Borin and R.J. Poppi, “Application of mid infrared spectroscopy and iPLS for quantification of contaminants in lubricating oil”, *Vibrational Spectroscopy*, 37 (Diciembre.2005).
55. C.-A. Baldrich Ferrer and L.-A. Novoa Mantilla, “Rapid characterization of diesel fuel by infrared spectroscopy”, *Ciencia, Tecnologia y Futuro (CT&F)*, **3**, 2 (2006).
56. R.M. Balabin, R.Z. Safieva and E.I. Lomakina, “Comparison of linear and nonlinear calibration models based on near infrared 5NIR) spectroscopy data for gasoline properties prediction”, *Chemometrics and Intelligent Laboratory Systems*, 88 (2007).
57. R.M. Balabin, R. Safieva and E.I. Lomakina, “Wavelet neural network (WNN) approach for calibration model building based on gasoline near infrared (NIR) spectra”, *Chemometrics and Intelligent Laboratory Systems*, 93 (2008).
58. P. Baptista, P. Felizardo, J.C. Menezes and 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).
59. D. Özdemir, “Near Infrared Spectroscopic Determination of Diesel Fuel Parameters Using Genetic Multivariate Calibration”, *Petroleum Science and Technology*, **26**, 1 (2008).
60. C.F. Pereira, M.F. Pimentel, R.K.H. Galvão, F.A. Honorato, L. Stragevitch and M.N. Martins, “A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers”, *Analytica Chimica Acta*, **611**, 1 (2008).
61. S. Amat-Tosello, N. Dupuy and J. Kister, “Contribution of external parameter orthogonalisation for calibration transfer in short waves--near infrared spectroscopy application to gasoline quality”, *Analytica Chimica Acta*, **642**, 1-2 (2009).
62. M.A. Al-Ghouti, Y. Al-Degs and M. Amer, “Application of chemometrics and FTIR for determination of viscosity index and base number of motor oils”, *Talanta*, 81 (2010).
63. R.M. Balabin, R.Z. Safieva and E.I. Lomakina, “Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines”, *Microchemical Journal*, 98 (2010).
64. A.A. Kardamakis and N. Pasadakis, “Autoregressive Modelling of near-IR spectra and MLR to predict RON values of gasolines”, *Fuel*, 89 (2010).
65. L. de Fátima Bezerra Lira, F.V.C. de Vasconcelos, C.F. Pereira, A.P.S. Paim, L. Stragevitch and M.F. Pimentel, “Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration”, *Fuel*, **89**, 2 (2010).
66. C.-A. Baldrich, L.-A. Novoa and A. Bueno, “Comparison between NIR and UVVIS spectra chemometrics for predicting FCC feedstocks properties”, *Ciencia Tecnologia y Futuro CT&F*, **4**, 1 (2010).
67. J.B. Cooper, C.M. Larkin and M.F. Abdelkader, “Virtual Standard Slope and Bias Calibration Transfer of Partial Least Squares Jet Fuel Property Models to Multiple near Infrared Spectroscopy Instruments”, *Journal of Near Infrared Spectroscopy*, **19**, 2 (2011).
68. R.M. Balabin, “Near Infrared (NIR) spectroscopy for biodiesel analysis: Fractional Composition, Iodine Value, and Cold Filter Plugging Point from One vibrational Spectrum”, *Energy & Fuels* (2011).
69. R.M. Balabin, E.I. Lomakina and 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**, 5 (2011).
70. S. Marinovic, A. Jukic, D. Dolezal, B. Speha and M. Kristovic, "Prediction of used lubrication oils properties by infrared spectroscopy using multivariate analysis", 3 (2012).
71. M. Bassbas, A. Hafid, S. Platikanov, R. Tauler and A. Oussama, "Study of motor oil adulteration by infrared spectroscopy and chemometrics methods", *Fuel*, 104 (2012).
72. Li Yan-kun, "Determination of diesel cetane number by consensus Modelling based on uninformative variable elimination", *Anal. Methods*, **4**, 1 (2012).
73. S. Marinovic, M. Kristovic, B. Spehar, V. Rukavina and A. Jukic, "Prediction of Diesel Fuel Properties by Vibrational Spectroscopy Using Multivariate Analysis", *Journal of Analytical Chemistry*, **67**, 12 (2012).
74. Z. Xu, *Prediction and classification of physical properties by Near-Infrared Spectroscopy and baseline correction of Gas Chromatogra^hy Mass Spectrometry data of Jet Fuels by using chemometrics algorithms*, Ohio (2012).
75. A. Villar, S. Fernández, E. Gorritxategi, J.I. Ciria and L.A. Fernández, "Optimization of the multivariate calibration of a Vis–NIR sensor for the on-line monitoring of marine diesel engine lubricating oil by variable selection methods", *Chemometrics and Intelligent Laboratory Systems*, 130 (2013).
76. C.R. Souza, A.H. Silva, N. Nagata, J.L.T. Ribas, F. Simonelli and A. Barison, "Cetane Number Assessment in Diesel Fuel by 1 H or Hydrogen Nuclear Magnetic Resonance-Based Multivariate Calibration", *Energy Fuels*, **28**, 8 (2014).
77. K. He, F. Qjan, H. Cheng and W. Du, "A novel adaptive algorithm with near-infrared spectroscopy and its application in online gasoline blending processes", *Chemometrics and Intelligent Laboratory Systems*, 140 (2015).
78. P.M. Santos, R.S. Amais, L.A. Colnago, Å. Rinnan and M.R. Monteiro, "Time Domain-NMR Combined with Chemometrics Analysis: An Alternative Tool for Monitoring Diesel Fuel Quality", *Energy Fuels*, **29**, 4 (2015).
79. F. Feng, Q. Wu and L. Zeng, "Rapid analysis of diesel fuel properties by near infrared reflectance spectra", *Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy*, **149** (2015).
80. A.G. Abdul Jameel, N. Naser, A.-H. Emwas, S. Dooley and S.M. Sarathy, "Predicting Fuel Ignition Quality Using 1 H NMR Spectroscopy and Multiple Linear Regression", *Energy Fuels*, **30**, 11 (2016).
81. Z.S. Baird and V. Oja, "Predicting fuel properties using chemometrics: a review and an extension to temperature dependent physical properties by using infrared spectroscopy to predict density", *Chemometrics and Intelligent Laboratory Systems*, **158** (2016).
82. C. Brouillette, W. Smith, C. Shende, Z. Gladding, S. Farquharson, R.E. Morris, J.A. Cramer and 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**, 5 (2016).
83. R.R. de Oliveira, R.H. Pedroza, A.O. Sousa, K.M. Lima and A. de. Juan, "Process Modelling and control applied to real-time monitoring of", *Analytica Chimica Acta* (2017).
84. A.B. Câmara, L.S. de Carvalho, C.L. de Morais, L.A. de Lima, K.O. de Araujo, F.M. de Oliveira and K.M. de Lima, "MCR-ALS and PLS coupled to NIR/MI spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends", *Fuel*, 210 (2017).
85. B. Leal de Rivas, J.-L. Vivancos, J. Ordieres-Meré and S.F. Capuz-Rizo, "Determination of the total acid number (TAN) of used mineral oils in aviation engines by FTIR using regression models", *Chemometrics and Intelligent Laboratory Systems*, 160 (2017).
86. A.S. Luna, I.C.A. Lima, C.A. Henriques, L.R.R. de Araujo, W. Da Fortunato Rocha and J.V. Da Silva, "Prediction of fatty methyl esters and physical properties of soybean oil/biodiesel blends from near and mid-infrared spectra using the data fusion strategy", *Anal. Methods*, **9**, 33 (2017).
87. R.R. Rodrigues, J.T. Rocha, L.M.S. Oliveira, J.C.M. Dias, E.I. Muller, V.E. Castro and P.R. Filgueiras, "Evaluation of calibration transfer methods using the ATR-FTIR technique to predict density crude oil", *Chemometrics and Intelligent Laboratory Systems*, 166 (2017).
88. B. Zhan and J. Yang, "Measurement of Diesel Cetane Number Using Near Infrared Spectra and Multivariate Calibration", *Advances in Engineering*, **100** (2017).
89. A. Palou, A. Miró, M. Blanco, R. Larraz, J.F. Gómez, T. Martínez, J.M. González and 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", *Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy*, **180** (2017).
90. N.C. da Silva, C.J. Cavalcanti, F.A. Honorato, J.M. Amigo and M.F. Pimentel, "Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters", *Analytica Chimica Acta*, **954** (2017).
91. J.-J. Da Costa Soares, *Compréhension moléculaire et prédiction des propriétés physico-chimiques dans les produits pétroliers*, Solaize-Lyon (2018).
92. M. Lacoue-Nègre, *Livrable commun PIW-L034 et PMC-L019. Performances des modèles de prédictions de propriétés et qualités produit de coupes à partir du spectre RMN 13C du liqtot*, Solaize (2017).
93. A.F. Constantino, D.C. Cubides-Román, R.B. dos Santos, L.H. Queiroz, L.A. Colnago, Á.C. Neto, L.L. Barbosa, W. Romão, E.V. de Castro, P.R. Filgueiras and V. Lacerda, "Determination of physicochemical properties of biodiesel and blends using low-field NMR and multivariate calibration", *Fuel*, **237** (2019).
94. I. Barra, M. Kharbach, E.M. Qannari, M. Hanafi, Y. Cherrah and A. Bouklouze, "Predicting cetane number in diesel fuels using FTIR spectroscopy and PLS regression", *Vibrational Spectroscopy*, **111** (2020).
95. A. Shukla, H. Bhatt and A.K. Pani, eds., *Variable selection and Modelling from NIR spectra data: A case study of diesel quality prediction using LASSO and Regression Tree* (2020).
96. P. Mishra, F. Marini, A. Biancolillo and J.-M. Roger, "Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques", *Talanta*, **223**, Pt 1 (2021).
97. I. Hradecká, R. Velvarská, K. Dlasková Jaklová and A. Vráblík, "Rapid determination of diesel fuel properties by near-infrared spectroscopy", *Infrared Physics & Technology* (2021).
98. H. Yu, X. Wang, F. Shen, J. Long and W. Du, "Novel automatic model construction method for the rapid characterization of petroleum properties from near-infrared spectroscopy", *Fuel*, **316** (2022).
99. S. Liu, S. Wang, C. Hu, X. Qin, J. Wang and D. Kong, "Development of a new NIR-machine learning approach for simultaneous detection of diesel various properties", *Measurement*, **187** (2022).
100. L.M. de Aguiar, D. Galvan, E. Bona, L.A. Colnago and 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).
101. B. Zhan and J. Yang, "Measurement of Diesel Cetane Number Using Near Infrared Spectra and Multivariate Calibration", in: *Proceedings of the 2017 International Conference on Manufacturing Engineering and Intelligent Materials (ICMEIM 2017),* Atlantis Press, Paris, France (25/02/2017 - 26/02/2017).
102. T. de Beer, A. Burggraeve, M. Fonteyne, L. Saerens, J. Remon and C. CVervaet, "Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes", *International Journal of Pharmaceutics*, 417 (2011).
103. H. Dalvi, A. Langlet, M.-J. Colbert, A. Cournoyer, J.-M. Guay, N. Abatzoglou and R. Gosselin, "In-line monitoring of Ibuprofen during and after tablet compression using near-infrared spectroscopy", *Talanta*, 195 (2019).
104. M. Verstraeten, D. van Hauwermeiren, M. Hellings, E. Hermans, J. Geens, C. Vervaet, I. Nopens and T. de Beer, "Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process", *International Journal of Pharmaceutics*, 546 (2018).
105. Intertek, "Crude Oil Analysis - INBLEND Real-Time Analysis" (2017).
106. J. Britain Cooper, "Chemometric analysis of Raman spectroscopic data for process control applications", *Chemometrics and Intelligent Laboratory Systems*, **46** (1999).
107. M. Drobon, J. Durand and Y. Boscher, "On-line octane-number analyser for reforming unit effluents . Principle of the analyser and test of a prototype", *Analytica Chimica Acta*, 238 (1990).
108. J. Durand, Y. Bischer and M. Dorbon, "On-line chromatographic analyser for determining the composition and octane number of reforming process effluents", *Journal of Chromatography*, 509 (1990).
109. P. Marteau, N. Zanier-Szydlowski, A. Aoufi, G. Hotier and C. François, "Remote Raman spectroscopy for process control", *Vibrational Spectroscopy* (1995).
110. P.P. Mortensen and R. Bro, "Real-time monitoring and chemical profiling of a cultivation process", *Chemometrics and Intelligent Laboratory Systems* (2006).

111. ASTM D 2892-20, “Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column)”. ASTM International, West Conshohocken, PA (2020).
112. M.A. Fahim, T.A. Alsahhaf and A. Elkilani, “Environmental Aspects in Refining”, in: *Fundamentals of Petroleum Refining,* Elsevier, pp. 423–455 (2010).
113. S. Wold, K. Esbensen and P. Geladi, “Principal component analysis”, *Chemometrics and Intelligent Laboratory Systems*, **2**, 1-3 (1987).
114. L. Rokach and O. Maimon, “Clustering Methods”. Springer-Verlag.
115. L.S. Chen, D. Paul, R.L. Prentice and P. Wang, “A regularized Hotelling's T2 test for pathway analysis in proteomic studies”, *Journal of the American Statistical Association*, **106**, 496 (2011).
116. J. Buendia Garcia, J. Gornay, M. Lacoue-Negre, S. Mas Garcia, J. Er-Rmyly, R. Bendoula and J.-M. Roger, “A novel methodology for determining effectiveness of preprocessing methods in reducing undesired spectral variability in near infrared spectra”, *Journal of Near Infrared Spectroscopy* (2022).
117. G. Tomasi, F. Savorani and S.B. Engelsen, “icoshift: An effective tool for the alignment of chromatographic data”, *Journal of chromatography. A*, **1218**, 43 (2011).
118. Abraham. Savitzky/M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.”, *Analytical Chemistry*, **36** (1964).
119. H. Martens, M. Høy, B.M. Wise, R. Bro and P.B. Brockhoff, “Pre-whitening of data by covariance-weighted pre-processing”, *Journal of Chemometrics*, **17**, 3 (2003).
120. G. Rabatel, F. Marini, B. Walczak and J.-M. Roger, “VSN: Variable sorting for normalization”, *Journal of Chemometrics*, **34**, 2 (2020).
121. R. J. Barnes, M. S. Dhanoa, & S. J. Lister, “Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra.”, *Applied Spectroscopy*, **43** (1989).
122. H. Martens and T. Naes, *Multivariate calibration*. Wiley, Chichester (1989).
123. F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, “Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics”, *Analytical Chemistry*, **78**, 13 (2006).
124. Å. Rinnan, F. den van Berg and S.B. Engelsen, “Review of the most common pre-processing techniques for near-infrared spectra”, *TrAC Trends in Analytical Chemistry*, **28**, 10 (2009).
125. M. Harald and Edward Stark, “Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy”, *Journal of Pharmaceutical & Biomedical Analysis*, **9** (1991).
126. K. H. Norris and P. C. Williams, “Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size.”, *Cereal Chemistry*, **61** (Mar. 1984).
127. A. Höskuldsson, “PLS regression methods”, *J. Chemometrics*, **2**, 3 (1988).
128. N. Cristianini and E. Ricci, “Support Vector Machines”, in: *Encyclopedia of algorithms,* Ed by M.-Y. Kao, Springer, New York (N.Y.), pp. 928–932 (2008).
129. S. Walczak and N. Cerpa, “Artificial Neural Networks”, in: *Encyclopedia of physical science and technology,* Ed by R.A. Meyers, Academic Press, San Diego, pp. 631–645 (2002).
130. T. Naes, T. Isaksson and B. Kowalski, “Locally weighted regression and scatter correction for near-infrared reflectance data”, *Anal. Chem.*, **62**, 7 (1990).
131. D.L. Hall and J. Llinas, “An introduction to multisensor data fusion”, *Proc. IEEE*, **85**, 1 (1997).
132. M. Cocchi, ed., *Data fusion methodology and applications*. Elsevier, Amsterdam (2019).
133. A. K. Smilde and I. V. Mechelen, “A Framework for Low-Level Data Fusion”, *Data Fusion Methodology and Applications* (2019).
134. A. Smolinska, J. Engel, E. Szymanska, L. Buydens and L. Blanchet, “General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences”, *Data Fusion Methodology and Applications* (2019).
135. D. Ballabio, R. Todeschini and V. Consonni, “Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data”, in: *Data fusion methodology and applications,* Ed by M. Cocchi, Elsevier, Amsterdam, pp. 129–155 (2019).
136. Q.C. Nguyen, K.H. Liland, O. Tomic, A. Tarrega, P. Varela and T. Næs, “SO-PLS as an alternative approach for handling multi-dimensionality in modelling different aspects of consumer expectations”,

*Food research international (Ottawa, Ont.)*, **133** (2020).
137. K.H. Liland, T. Naes and U.G. Indahl, "ROSA-a fast extension of partial least squares regression for multiblock data analysis", *J. Chemometrics*, **30**, 11 (2016).
138. S. Wold, A. Johansson and M. Cochi, *PLS-partial least squares projections to latent structures.* (1993).
139. T. Rajalahti, R. Arneberg, F.S. Berven, K.-M. Myhr, R.J. Ulvik and O.M. Kvalheim, "Biomarker discovery in mass spectral profiles by means of selectivity ratio plot", *Chemometrics and Intelligent Laboratory Systems*, **95**, 1 (2009).
140. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso.", *Journal of the Royal Statistical Society. Series B (Methodological)*, **58** (1996).
141. D.E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley publishing company, Reading (Mass.) etc. (1989).
142. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", **46** (2002).
143. M. Last, A. Kandel and O. Maimon, "Information-theoretic algorithm for feature selection", *Pattern Recognition Letters*, **22**, 6-7 (2001).
144. S. Nakariyakul and D.P. Casasent, "An improvement on floating search algorithms for feature subset selection", *Pattern Recognition*, **42**, 9 (2009).
145. J.M. Roger, B. Palagos, D. Bertrand and E. Fernandez-Ahumada, "CovSel: Variable selection for highly multivariate and multi-response calibration", *Chemometrics and Intelligent Laboratory Systems*, **106**, 2 (2011).
146. A. Biancolillo, F. Marini and J.-M. Roger, "SO-CovSel: A novel method for variable selection in a multiblock framework", *J. Chemometrics*, **34**, 2 (2020).
147. L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck and S.B. Engelsen, "Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy", *Appl Spectrosc*, **54**, 3 (2000).
148. Å. Rinnan, M. Andersson, C. Ridder and S.B. Engelsen, "Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS", *J. Chemometrics*, **28**, 5 (2014).
149. F. Chauchard, J.M. Roger and V. Bellon-Maurel, "Correction of the temperature effect on near infrared calibration—application to soluble solid content prediction", *Journal of Near Infrared Spectroscopy*, **12** (2004).
150. Y. Wang, D.J. Veltkamp and B.R. Kowalski, "Multivariate instrument standardization", *Anal. Chem.*, **63**, 23 (2002).
151. R.N. Feudale, N.A. Woody, H. Tan, A.J. Myles, S.D. Brown and J. Ferré, "Transfer of multivariate calibration models: A review", *Chemometrics and Intelligent Laboratory Systems*, **64**, 2 (2002).
152. J.-M. Roger, F. Chauchard and V. Bellon-Maurel, "EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits", *Chemometrics and Intelligent Laboratory Systems*, **66**, 2 (2003).
153. M. Zeaiter, J.M. Roger and 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", *Chemometrics intelligent laboratory systems* (2005).
154. J. Buendia Garcia, M. Lacoue-Negre, J. Gornay, S. Mas Garcia, R. Bendoula and J.M. Roger, "Diesel cetane number estimation from NIR spectra of hydrocracking total effluent", *Submitted to Fuel* (2022).
155. E.D. Yalvac, M.B. Seasholtz and S.R. Crouch, "Evaluation of Fourier Transform Near-Infrared for the Simultaneous Analysis of Light Alkene Mixtures", *Appl. Spectrosc., AS*, **51**, 9 (1997).
156. J.J. Kelly and J.B. Callis, "Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines", *Anal. Chem.*, **62**, 14 (1990).
157. M. Zeaiter, J.-M. Roger, V. Bellon-Maurel and D.N. Rutledge, "Robustness of models developed by multivariate calibration. Part I", *TrAC Trends in Analytical Chemistry*, **23**, 2 (2004).
158. R. Legner, M. Voigt, A. Wirtz, A. Friesen, S. Haefner and M. Jaeger, "Using Compact Proton Nuclear Magnetic Resonance at 80 MHz and Vibrational Spectroscopies and Data Fusion for Research Octane Number and Gasoline Additive Determination", *Energy Fuels*, **34**, 1 (2020).
159. T.I. Dearing, W.J. Thompson, C.E. Rechsteiner and B.J. Marquardt, "Characterization of Crude Oil Products Using Data Fusion of Process Raman, Infrared, and Nuclear Magnetic Resonance (NMR)

Spectra", *Appl Spectrosc*, **65**, 2 (2011).
160. M.K. Moro, Á.C. Neto, V. Lacerda, W. Romão, L.S. Chinelatto, E.V. Castro and P.R. Filgueiras, "FTIR, 1H and 13C NMR data fusion to predict crude oils properties", *Fuel*, **263** (2020).
161. M. Zeaiter, J.-M. Roger and V. Bellon-Maurel, "Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods", *TrAC Trends in Analytical Chemistry*, **24**, 5 (2005).
162. R. J. Barnes, M. S. Dhanoa and Susan J. Lister, "Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra", *Appl Spectrosc*, **43**, 5 (1989).
163. M. Harald and Edward Stark, "Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy", *Journal of Pharmaceutical & Biomedical Analysis*, **9** (1991).
164. A. Rinnan, F. van den Berg and S.B. Egelsen, "Review of the most common pre-processing techniques for near-infrared spectra", *Trends in analytical chemistry*, **25**, 10 (2009).
165. S. Wold, H. Antti, F. Lindgren and J. Ohman, "Orthogonal signal correction of near-infrared spectra", *Chemometrics and Intelligent Laboratory Systems*, 44 (1998).
166. J. Engel, J. Gerretzen, E. Szymanska, J.J. Jansen, G. Downey, L. Blanchet and L.M. Buydens, "Breaking with trends in pre-processing", *Trends in analytical chemistry*, 50 (2013).
167. Z. Chen, D. Lovett and J. Morris, "Process analytical technologies and real time process control a review of some spectroscopic issues and challenges", *Journal of Process Control*, 21 (2011).
168. Z.-P. Chen, J. Morris and E. Martin, "Correction of temperature-induced spectral variations by loading space standardization", *Anal. Chem.*, **77**, 5 (2005).
169. Z.-P. Chen, J. Morris and E. Martin, "Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction", *Anal. Chem.*, **78**, 22 (2006).
170. I. Markovsky and S. van Huffel, "Overview of total least-squares methods", *Signal Processing*, **87**, 10 (2007).
171. L.E. Eberly, "Multiple linear regression", *Methods in molecular biology (Clifton, N.J.)*, **404** (2007).
172. D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning representations by back-propagating errors", *Nature*, **323**, 6088 (1986).
173. D.F. Specht, "Probabilistic neural networks", *Neural Networks*, **3**, 1 (1990).
174. L. Breiman, *Mach Learn*, **45**, 1 (2001).
175. F. Ferraty, A. Goia, E. Salinelli and P. Vieu, "Functional projection pursuit regression", *TEST*, **22**, 2 (2013).
176. D. Lahat, T. Adah and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects" (2015).
177. M. Cocchi and Elsevier, eds., *Data Fusion Methodology and Applications*. Elsevier (2019).
178. E. Borras, J. Ferré, R. Boqué, M. Mestres, L. Aceña, A. Calvo and O. Busto, "Prediction of olive oil sensory descriptors using instrumental data fusion and partial least squares(PLS) regression", *Talanta*, **155** (2016).
179. F. Comino, M.J. Ayora-Cañada, V. Aranda, A. Diaz and A. Dominguez-Vidal, "Near-infrared spectroscopy and X-ray fluorescence data fusion for olive leaf analysis and crop nutritional status determination", *Talanta*, **188** (2018).
180. M.S. Godinho, M.R. Blanco, F.F. Gambarra Neto, L.M. Liao, M.M. Sena, R. Tauler and A.E. de Oliveira, "Evaluation of transformer insulating oil quality using NIR, fluorescence, and NMR spectroscopic data fusion", *Talanta*, **129** (2014).
181. C. Márquez, M.I. López, I. Ruisánchez and M.P. Callao, "FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud", *Talanta*, **161** (2016).
182. S. Mas, R. Tauler and A. de Juan, "Chromatographic and spectroscopic data fusion analysis for interpretation of photodegradation processes", *Journal of Chromatography A*, **1218** (2011).
183. P.M. Ramos, I. Ruisánchez and K.S. Andrikopoulos, "Micro-Raman and X-ray fluorescence spectroscopy data fusion for the classification of ochre pigments", *Talanta*, **75** (2008).
184. J. Esteban, A. Starr, R. Willetts, P. Hannah and P. Bryasnton-Cross, "A Review of Data Fusion Models and Architectures: Towards Engineering Guidelines", *Neural computing & applications*, 14 (2004).

185. W. Saeys, N. Nguyen Do Trong, R. van Beers and B. M. Nicolai, “Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review”, *Postharvest Biology and Technology*, 158 (2019).
186. EigenVector, “wiki eigenvector”, http://wiki.eigenvector.com/index.php?title=T-Squared_Q_residuals_and_Contributions (2012).
187. D. Cousineau and S. Chartier, “Outliers detection and treatment: A review”, *International Journal of Psychological Research*, **3**, 1 (2010).
188. C.M. Andersen and R. Bro, “Variable selection in regression-a tutorial”, *Journal of Chemometrics*, **24**, 11-12 (2010).
189. M.J. Anzanello and F.S. Fogliatto, “A review of recent variable selection methods in industrial and chemometrics applications”, *EJIE*, **8**, 5 (2014).
190. G. Heinze, C. Wallisch and D. Dunkler, “Variable selection - A review and recommendations for the practicing statistician”, *Biometrical journal. Biometrische Zeitschrift*, **60**, 3 (2018).
191. W. Du, Z.-P. Chen, L.-J. Zhong, S.-X. Wang, R.-Q. Yu, A. Nordon, D. Littlejohn and M. Holden, “Maintaining the predictive abilities of multivariate calibration models by spectral space transformation”, *Analytica Chimica Acta*, **690**, 1 (2011).
192. E.J. Bjerrum, M. Glahder and T. Skov, *Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics* (05/10/2017).

# Appendices

# Appendix 1: Publication #1. “A novel methodology for determining effectiveness of preprocessing methods in reducing undesired spectral variability in near infrared spectra”

*Original Research Article*

JNIRS

Journal of Near Infrared Spectroscopy
2022, Vol. 0(0) 1–15

DOI: 10.1177/09670335211047959
journals.sagepub.com/home/jns
SAGE

# A novel methodology for determining effectiveness of preprocessing methods in reducing undesired spectral variability in near infrared spectra

Jhon Buendia Garcia[1,2], Julien Gornay[1], Marion Lacoue-Negre[1], Silvia Mas Garcia[3,2], Jihane Er-Rmyly[1], Ryad Bendoula[3] and Jean-Michel Roger[3,2]

## Abstract

This study uses a novel analysis methodology based on the Hierarchical Clustering Analysis (HCA) to determine the effectiveness of different preprocessing methods in minimizing undesired spectral variability in near infrared spectroscopy due to both the consecutive and repetitive acquisition of the spectrum and the sample temperature. Nine preprocessing methods and different combinations of them were evaluated in four case studies: reproducibility, repeatability, sample temperature, and combination of the before mentioned cases. Eighty-four spectra acquired on seven different hydrocarbon samples from catalytic conversion processes have been selected as the real case study to illustrate the potential of the mentioned methodology. The approach proposed allows a more detailed discriminatory analysis compared to the classical methods for comparing the between-class and the within-class variances, such as the Wilks' lambda criterion, and hence constitutes a powerful tool to determine adequate spectral preprocessing strategies. This study also proves the potential of the discrimination analysis methodology as a general scheme to identify atypical behaviors either in the spectrum acquisition or in the measured samples.

## Keywords

Spectral variability, hierarchical clustering analysis, principal components analysis, preprocessing effectiveness, outliers, Qresidual, Hotelling's $T^2$, reproducibility, repeatability, sample temperature

Received 4 June 2021; accepted 29 August 2021

## Introduction

In the past few decades, the use of near infrared (NIR) spectroscopy in the development of non-destructive and rapid measurement applications has been significantly increasing in several industries, such as food,[1,2] pharmaceuticals,[3,4] and petroleum.[5,6] Due to the recent growth boom in using NIR spectroscopy for real-time acquisition data,[7,8] the need to determine, analyze, and minimize spectral variability that is not associated with the physicochemical characteristics of the sample has notably arisen. This need becomes particularly evident when spectral variability is mainly generated by factors associated with the spectrum acquisition, such as spectrometer system, operator, measurement conditions, and environmental factors such as temperature and humidity, rather than the physicochemical characteristics of the sample. An example of this is the possible generation of spectral variability in the consecutive and repetitive acquisition of NIR spectra on a sample whose physicochemical characteristics remain constant over the spectrum acquisition. A lack of minimization of this type of spectral variability can result in inaccurate analysis and interpretation of spectroscopic information, misleading conclusions and flawed decision making.[9,10]

Among the classical performance parameters needed to validate a measurement methodology, precision is the most affected by the aforementioned factors. Precision is defined as the closeness of agreement between measured values obtained by replicate measurements on the same or similar samples under conditions of repeatability or reproducibility.[11] Repeatability conditions include the same measurement procedure, the same operator, the same instrument and measurement conditions, the same location, and a short interval between repetitions.[12] On the other hand, reproducibility implies successive measurements of the same sample under

[1]IFP Energies Nouvelles, Rueil-Malmaison, France
[2]ChemHouse Research Group, Montpellier, France
[3]ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France

Corresponding author:
Jhon Buendia Garcia, IFP Energies nouvelles Solaize, Rond-point de l'échangeur de Solaize, Solaize 69360, France.
Email: jhon.andersson@gmail.com

changing measurement conditions,[13] such as measurement principle, measurement method, operator, measurement instrument, reference standard, location, conditions of use, and time. spectral acquisition is very sensitive to any change in measurement. Despite ensuring that both the spectral acquisition conditions and the physicochemical characteristics of the sample do not change in a repetitive NIR spectral acquisition, the resulting spectra may have differences that can lead to random errors and deviations, which must be corrected or minimized.

Due to its high impact on NIR spectral acquisition accuracy, temperature is the most studied influencing parameter.[14,15] Hansen et al.[16] showed that molecular bond vibration intensity depends on temperature, leading to changes in the spectrum according to temperature variation. Furthermore, some physicochemical properties of samples, such as viscosity and density, are temperature-dependent, and these changes in the sample due to temperature are not permanent and do not reflect the intrinsic nature of the sample.[17–19] Nevertheless, these changes can significantly affect spectral acquisition. As with the spectral variability generated by repetitive spectrum acquisition of a specific sample, the variability caused by sample temperature must be minimized to ensure the reliable description of the sample physicochemical behavior from the spectroscopic information extracted.

Data preprocessing is a common step for reducing undesired effects and for minimizing spectral variability. There are different preprocessing algorithms for the correction of the undesired spectral variation; these can be divided into two main categories: scatter-correction methods, employed to correct the additive and multiplicative effects, and spectral derivatives, used to minimize the sources of unwanted and non-informative spectral variations.[20] Among the most common preprocessing methods used in NIR spectroscopy, Savitsky–Golay derivative (Sav–Gol),[21] Extended Multiple Signal Correction (EMSC),[22] Standard Normal Variate (SNV),[23] and recently, Variable Sorting for Normalization (VSN),[24] can be highlighted. However, the effectiveness of preprocessing methods is highly dependent on the type of spectroscopic information analyzed and the factors that are causing its variability.[9]

Evaluation of the preprocessing method effectiveness is generally based on the performance of prediction models.[25–27] Among the contributions reported in the literature, the work of Gerretzen et al.,[28] which presents a novel approach for the selection of the most appropriate preprocessing methods based on the design of experiments, is worth mentioning. Similarly, the studies of Devos et al.[29] and Allegrini et al.[30] which, by means of a parallel workflow approach of preprocessing and variable selection, present an interesting alternative to the optimization of the preprocessing method selection. Nonetheless, the application of these approaches may be limited when the variability of the physicochemical characteristics of the samples is negligible, but significant spectral variability exists as a result of the repetitive spectrum acquisition and the sample temperature. In that case, a different analysis approach may yield more detailed results, helping to improve understanding of the impact of these parameters. Another less common approach to assessing preprocessing methods effectiveness is analyzing the spectral variance.[31] Different statistical tools are available to determine both within-class variance (multiple measurements of the same sample) and between-class variance (measurements of different samples). One of the most common criteria used to evaluate between-class and within-class variances is the Wilks' lambda.[32]

In this study, a novel and general strategy based on Hierarchical Clustering Analysis (HCA)[33] was proposed for evaluating the effectiveness of preprocessing methods in reducing the spectral variability generated by parameters related to the continuous and dynamic spectrum acquisition. To this aim, the effectiveness of nine preprocessing methods and different combinations of them in minimizing undesired spectral variability due to repeatability, reproducibility, sample temperature, and combination of these parameters was evaluated. Eighty-four spectra acquired on seven different hydrocarbon samples from catalytic conversion processes have been selected as the real case study to illustrate the potential of the mentioned methodology.

To obtain reliable conclusions and validate the results obtained by the analysis methodology proposed, the Wilks' lambda criterion[32] was used as a reference method.

## Material and methods

### Samples

Twenty-four vacuum gasoil (VGO) samples were processed in the catalytic conversion pilot plant reactors at IFPEN (Solaize, France). From these reactors, ninety-three different hydrocarbon samples, known as total effluent, were obtained (see references[34,35] for a detailed description of catalytic conversion processes). From these 93 samples, 7 samples were selected, ensuring their representativeness and physicochemical diversity. Table 1 summarizes four relevant physicochemical properties of the selected samples: the density[36], the simulated initial boiling point and the distillation temperature range to obtain both 5% and 95% of sample distillate (Simulated Distillation IBP, T5 and T95).[37] It can be observed that physicochemical variability between the selected samples is guaranteed.

### Spectral acquisition

The spectra were recorded with a Fourier-transform nearinfrared (FT-NIR) spectrometer (Matrix-F, Bruker Optik GmbH, Ettlingen, Germany) within the range of 9090–4600 $cm^{-1}$ and a resolution of 4 $cm^{-1}$. A total of 32 scans were used to obtain the final spectrum for each measurement. For acquiring absorbance spectra, the spectrometer system was equipped with an immersion transflectance probe with an optical path fixed at 2 mm withstanding temperatures ranging from −40°C to 200°C. The software used with the spectrometer was OVP (OPUS Validation Program—Bruker Optik GmbH, Ettlingen, Germany) which automatically performs a series of analyses of the instrument's performance, evaluates them, and ensures that it is operating within specifications. In addition, to ensure the spectrometer operation within specifications and that the spectral variability generated was due to the parameters evaluated and not to the instrument's

Table 1. Samples physicochemical properties. SimDis IBP, T5 & T95 description: Simulated distillation to determine the temperatures to start the sample evaporation and to recover both 5% and 95% of sample distillate.

| Sample physicochemical properties | | | | |
|---|---|---|---|---|
| | | Distillation temperatures | | |
| Sample ID | Density g/mL | IBP (°C) | SimDis T5 (°C) | SimDis T95 (°C) |
| Sample 1 | 0.8049 | 84.7 | 121.6 | 425.8 |
| Sample 2 | 0.8186 | 79.8 | 111.5 | 474.5 |
| Sample 3 | 0.8219 | 80.3 | 117.8 | 516.0 |
| Sample 4 | 0.8411 | 83.8 | 136.1 | 503.5 |
| Sample 5 | 0.8470 | 83.7 | 129.1 | 512.6 |
| Sample 6 | 0.8962 | 148.9 | 227.6 | 504.6 |
| Sample 7 | 0.9181 | 159.8 | 231.6 | 502.9 |

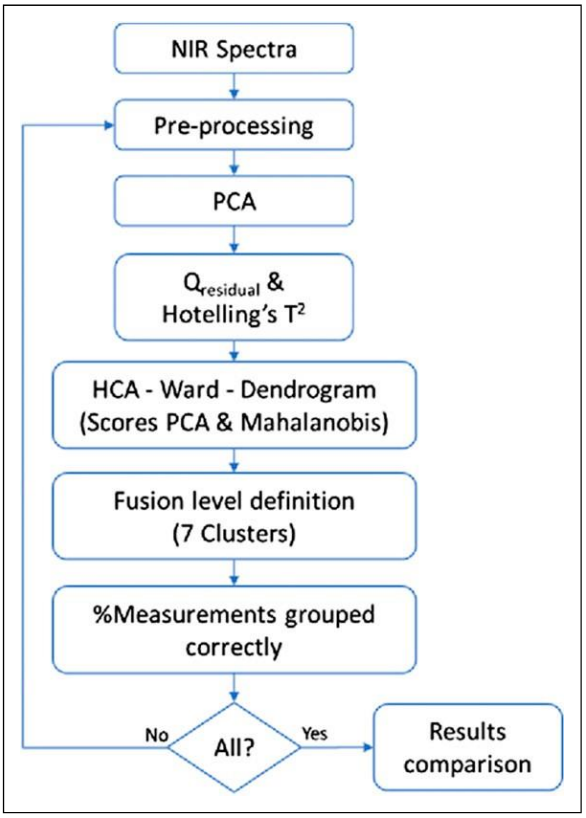


Figure 1. Methodology flow diagram.

inadequate functioning, the spectrometer performance was validated once a day using cyclohexane as an external reference sample. Before NIR analysis, the samples were heated in closed flasks at 60°C for 1 h in a water bath and shaken manually to ensure their liquid state and homogeneity. The initial boiling point (IBP) reported in Table 1 guarantees no loss of volatiles.

Ensuring the integrity and stability of both the sample and the NIR spectrum acquisition conditions, spectral variability due to repeatability, reproducibility, and sample temperature was generated. A short description of the spectrum acquisition for the cases evaluated in this study is presented below.

- Case 1. Spectral variability due to reproducibility: Each of the seven samples was analyzed once per day for five consecutive days at 60°C. Thirty-five spectra were obtained.
- Case 2. Spectral variability due to repeatability: Each of the seven samples was analyzed three times on the same day at 60°C. All samples were analyzed in less than 8 h. Twenty-one spectra were obtained.
- Case 3. Spectral variability due to the sample temperature: Each of the seven samples was analyzed at five different temperatures, ranging from 60°C to 80°C with a temperature increment of 5°C. The samples were heated in closed flasks at the desired temperature for 1 h. Evaporation losses of volatiles were null or negligible (see IBP in Table 1). Twenty-eight spectra were obtained.
- Case 4. Spectral variability due to the combination of the aforementioned cases: In this case, all spectra acquired in the above-described cases were used.

For each case, a matrix was generated. Each analyzed sample was defined as a class; thus, seven classes were defined in all matrices.

## Analysis methodology

The main steps of the data analysis workflow proposed in this study are schematized in Figure 1. A brief explanation of the procedure is given as follows.

The first step consisted in preprocessing each of the generated matrices. The nine most common preprocessing methods used in NIR data were divided into two categories: filtering and normalization methods. The preprocessing methods from each category were analyzed individually. If the total reduction or compensation of the studied spectral variability was not achieved, the evaluated preprocessing method was complemented with the methods belonging to the opposite category. This allowed the evaluation of all possible combinations and order of use of the preprocessing methods. The methods evaluated are described in Table 2. It should be emphasized that each preprocessing scenario evaluated includes the data centering by columns.

Afterwards, a preliminary inspection and dimension reduction of corresponding dataset were performed by using principal component analysis (PCA).[41] The number of chosen principal components (PCs) captured at least 99 % of the total variance in the dataset. The Q residual and Hotelling's $T^2$ tests were performed to determine the possible presence of anomalous data.[42]

Table 2. Preprocessing methods description.

| # | Category | Name | Acronym | Parameters |
|---|---|---|---|---|
| 1 | Filtering | Automatic Weighted Least Squares Baseline[20] | AWLS-B | |
| 2 | | Norris-Williams Derivation[38] | NW-D | 15-Point window, gap size = 7, First-order derivation |
| 3 | | Savitsky–Golay Derivative[21] | SG-D | 15-Point window, polynomial order = 2 first-order derivative |
| 4 | | Detrend[23] | Dtd | Polynomial order = 1 |
| 5 | | Extended Multiplicative Scatter/ Signal Correction[22] | EMSC | Reference spectrum (basis to remove the scatter) = mean of each matrix generated, polynomial order = 2, whole spectral range |
| 6 | Normalization | Multiplicative Signal Correction[39] | MSC | Reference data = mean of data, whole spectral range |
| 7 | | Standard Normal Variate[23] | SNV | |
| 8 | | Probabilistic Quotient Normalization[40] | PQN | |
| 9 | | Variable Sorting for Normalization[24] | VSN | Automatic calculation |

The chosen PCs scores were then used to perform the hierarchical clustering analysis (HCA) employing Ward's algorithm and Mahalanobis distance. The HCA aims to group clusters to form a new one to either minimize a statistical distance between classes or maximize a measure of similarity between them.[43,44] The analysis starts with as many groups as individuals contained in the dataset. From these initial groups, clusters are formed in an ascending manner until all cases treated are included in at least one of them. Ward's algorithm seeks to minimize each group variance by calculating all samples mean in each cluster. The algorithm then calculates each case distance and the cluster mean, adding up the distances between all cases. Finally, the clusters whose sum of distances is minimal are grouped. This procedure creates homogeneous groups of a similar amount of individuals. For achieving the grouping of classes, it is necessary to define a comparison parameter to calculate the variance of each class concerning the others. The most common is the Mahalanobis distance.[39]

A common manner of displaying the cluster analysis results is constructing a tree diagram known as a dendrogram. The resulting diagram shows the different groups' clustering order and the association measure's value, also known as the fusion level. The fusion level was defined for obtaining seven clusters corresponding to the seven classes. Finally, the number of correctly grouped sample measurements in each cluster was determined, and the percentage of clustering was calculated as the following: Number of correctly grouped samples/Total number of samples * 100.

These steps were repeated for each preprocessing method scenario, and the results obtained were compared using the percentage of samples correctly grouped as a figure of merit to determine the effectiveness of the preprocessing methods evaluated.

All the analyses were conducted with the PLS_Toolbox version 8.8 (Eigenvector Research Inc., Wenatchee, WA, USA) for MATLAB version R2019b (MathWorks, Natick, MA, USA).

### *Results comparison*

To validate the results obtained with the methodology proposed in this study, the Wilks' lambda was used as a comparative criterion. This criterion evaluates how well the data set classes are separated by calculating a ratio involving between-class and within-class variances. Several versions of the Wilks' lambda exist. In this article, the ratio of the between-class variance over the total variance was used. This ratio varies between 0 and 1, where 0 means that all the classes are superimposed, and 1 means that all the classes are perfectly separated.

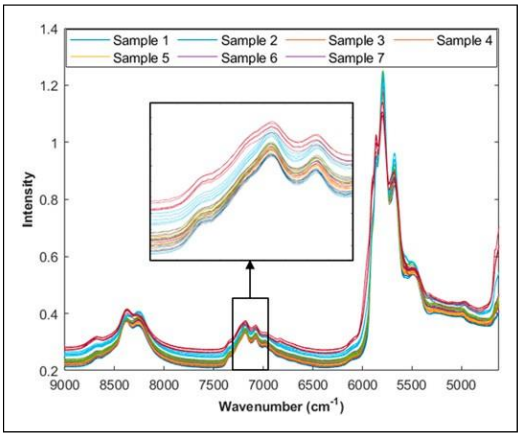



Figure 2. 35 raw spectra used in Case 1 evaluation over the entire spectral range. Legend: Color → samples analyzed. Black square → 35 raw spectra used in Case 1 evaluation magnified over the 7463 $cm^{-1}$ - 6993 $cm^{-1}$ spectral range.

## Results and discussion

This section shows the results obtained from the application of the proposed methodology. A comprehensive description of its application for the case study of reproducibility (Case 1) is presented. However, only the main results from the case studies of repeatability, temperature effect, and the combination of all cases are showed. Finally, the validation and the advantages of the proposed methodology are offered.

### *Case 1—Reproducibility*

For the reproducibility case, 35 spectra were used (5 daily spectra for each sample, Figure 2 shows the raw spectra over the entire spectral range used. At first glance, it can be

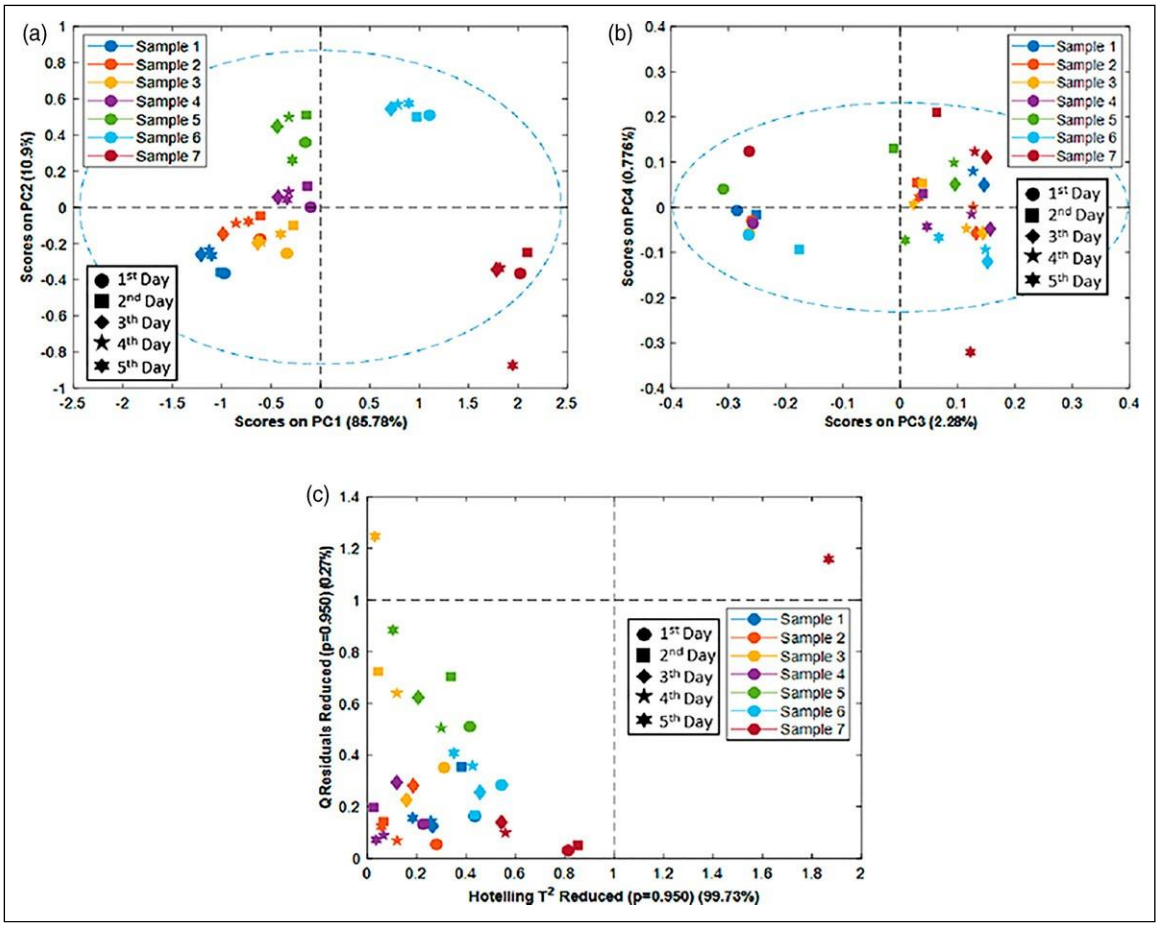


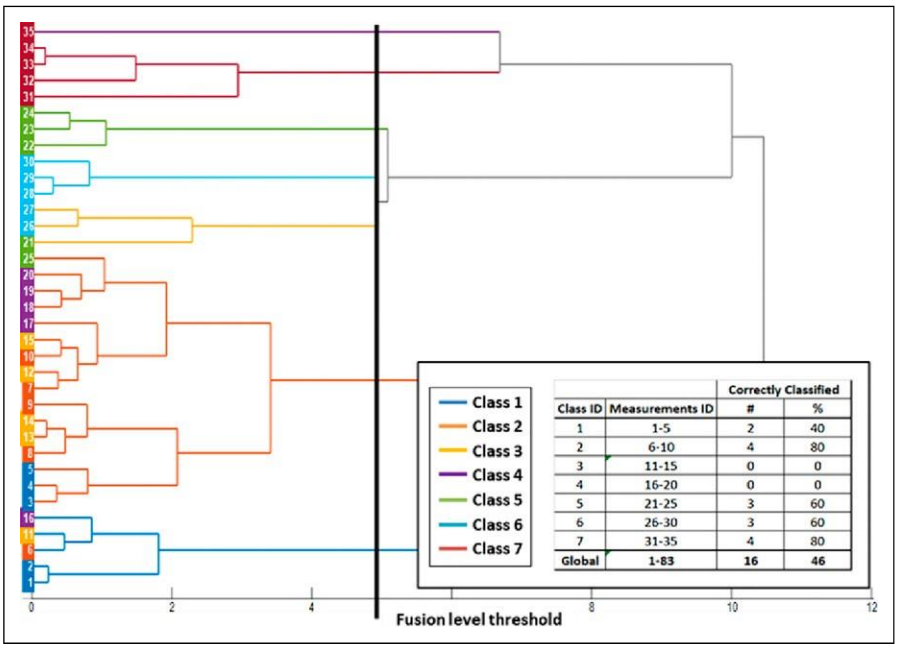

Figure 3. (a) Score plot of PC1 and PC2 for Case 1 with centered spectra, (b) score plot of PC3 and PC4 for Case 1 with centered spectra, and (c) reduced Qresidual and Hotelling's $T^2$ for Case 1 using 4 PCs with centered spectra. Legend: Shapes → acquisition day. Color → samples analyzed (classes).

observed the variability of spectra due to the physicochemical nature of the sample, showing, with some exceptions, a trend consistent with the properties reported in Table 1; that is, the spectra acquired on the sample with the lowest density (sample 1—dark blue) is at the bottom of the plot, and the spectra acquired on the sample with the highest density (sample 7—red) is at the top. This observation becomes more evident in the black square of Figure 2 where the spectra are magnified over a defined range 7463 $cm^{-1}$-6993 $cm^{-1}$. In this same figure, it is possible to visualize that the reproducibility measurements made on a single sample generate a variability that has a similar behavior to the variability caused by multiplicative effects, which can impact the final results obtained.

In this first case, the proposed methodology application shows the impact of the variability in NIR spectra acquisition caused by reproducibility measurements as well as the effectiveness of preprocessing methods in minimizing this spectral variability.

First, an exploratory analysis using PCA was performed to visualize the similarity/dissimilarity among the 5 spectra (1 per day) acquired for each of the 7 hydrocarbon samples analyzed. The data set (35 spectra) was merely centered. The first two components (PC1 and PC2) explained 96.7% of the data set total variance. From the score plot of the first two components (see Figure 3(a)), it can be seen that all 5 measurements of samples 1 to 6 are grouped in a consistent pattern regarding to the variance within classes. However, there are measurements of different samples that intersect with each other (between-class variance) (see Figure 3(a) and (b)), and hence their clustering can be influenced. Additionally, from these score plots, it can also be observed that the measurement of sample 7 on the fifth day is relatively distant from the other measurements of this sample. This could mean the presence of possible outliers in the dataset. Figure 3(c) shows the Hotelling's $T^2$ and Q residual scores for the dataset. It can be observed that the same measurement identified previously (Sample 7, fifth acquisition day) was found to be above the threshold of both tests. Therefore, this measurement can be confirmed as an outlier.

The scores of the first 4 principal components yielded by the PCA analysis (99% explained variance) were used to perform the HCA analysis. Figure 4 shows the dendrogram obtained from the HCA, where a fusion level (black line) for obtaining seven clusters corresponding to the seven hydrocarbon samples is defined. This Figure also shows the correct grouping percentage achieved for each class. From the table embedded in the Figure, it can be observed that no sample achieves an accurate grouping of all its measurements. Samples 2 and 7 (orange and red classes) are the classes with the highest correct grouping percentage (4 out of 5 for 80%), while classes 3 and 4 (yellow and purple classes) do not have any correctly grouped measurements.




| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-5 | 2 | 40 |
| 2 | 6-10 | 4 | 80 |
| 3 | 11-15 | 0 | 0 |
| 4 | 16-20 | 0 | 0 |
| 5 | 21-25 | 3 | 60 |
| 6 | 26-30 | 3 | 60 |
| 7 | 31-35 | 4 | 80 |
| Global | 1-83 | 16 | 46 |

Figure 4. Dendrogram for Case 1 (all measurements) with centered spectra. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class.

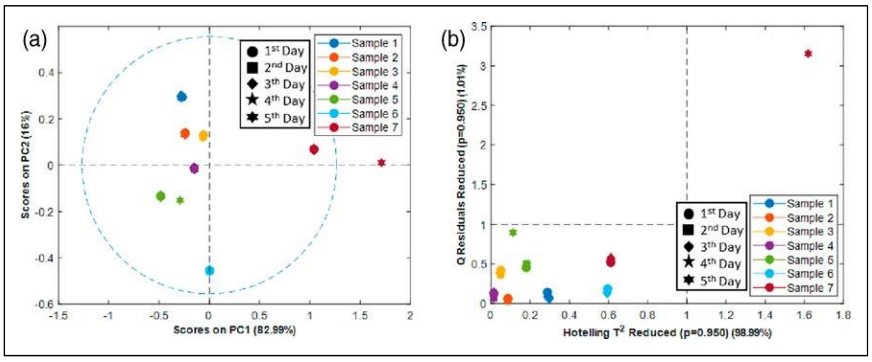
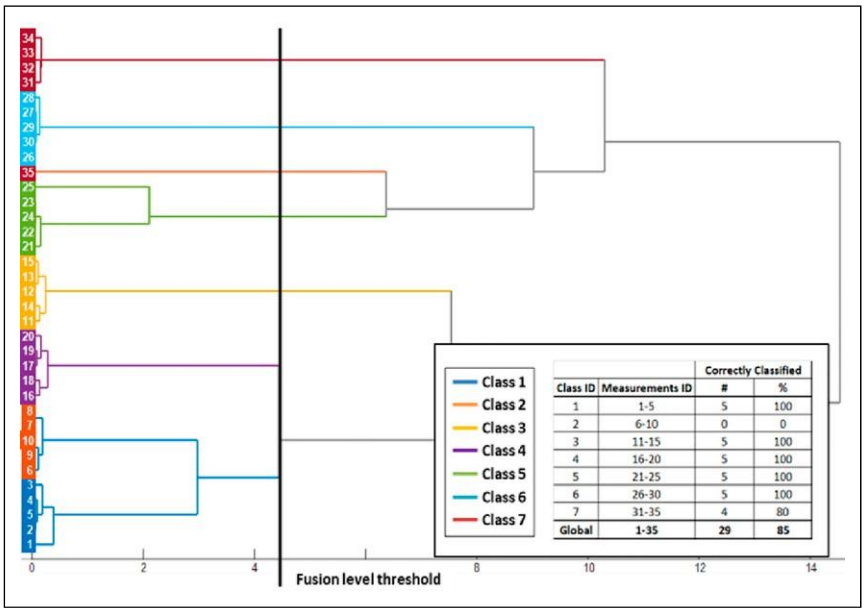



Figure 5. (a) Score plot of PC1 and PC2 for Case 1 with EMSC preprocessing and (b) reduced Qresidual and Hotelling's $T^2$ for Case 1 using 2 PCs with EMSC preprocessing. Legend: Shapes → day of measurement. Color → samples measured (classes).

Furthermore, measurement 35 that belongs to class 7 (red class) and identified in the previous steps as a potential outlier is the only measurement grouped in class 4 (purple class). Based on these results, it can be considered that this measurement has no similarity with any of the other 34 measurements, which confirms its outlier status.

In order to compensate the spectral variability caused by the lack of reproducibility and hence achieve a better clustering of all the classes, the use of appropriate preprocessing methods is needed. Nine preprocessing methods and different combinations of them were evaluated. Figure A1, which can be found in the article's supplementary section, summarizes the correct grouping percentage results for each preprocessing scenario evaluated in the 4 cases. From this Figure, it can be seen that for Case 1 (blue diamond), none of the evaluated scenarios reaches a correct grouping percentage of 100%, implying that none of the scenarios achieved the total reduction of the spectral variability generated by the reproducibility measurements. With a correct grouping of 85%, the EMSC was the most effective preprocessing method scenario. Figure 5 shows for this scenario the score plot of the first two principal components of the PCA analysis and the Q residual and Hotelling's $T^2$ results achieved using 2 PCs. It can be seen that all measurements are consistently grouped and are within the threshold of the two tests, except for measurement 35 (potential outlier identified). From the HCA dendrogram and the table of its corresponding correct grouping percentage (see Figure 6), it could be observed that this measurement is still incorrectly grouped as the unique measurement in class 2 (orange class). From these results, it can be assumed that all other measurements could be correctly grouped without this misgrouped measurement. Therefore, measurement 35 was removed from the data set, and the preprocessing method scenarios were re-evaluated to confirm this assumption.

Figure A1 (red diamond) shows that by removing measurement 35 from the data set, correct grouping of all measurements is possible in 6 scenarios (EMSC, AWLS-B+MSC, AWLS-B+SNV, AWLS-B+VSN, SG-D+SNV, and MSC+AWLS-B). However, only one scenario uses a single




| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-5 | 5 | 100 |
| 2 | 6-10 | 0 | 0 |
| 3 | 11-15 | 5 | 100 |
| 4 | 16-20 | 5 | 100 |
| 5 | 21-25 | 5 | 100 |
| 6 | 26-30 | 5 | 100 |
| 7 | 31-35 | 4 | 80 |
| Global | 1-35 | 29 | 85 |

Figure 6. Dendrogram for Case 1 (all measures) with EMSC preprocessing. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class.

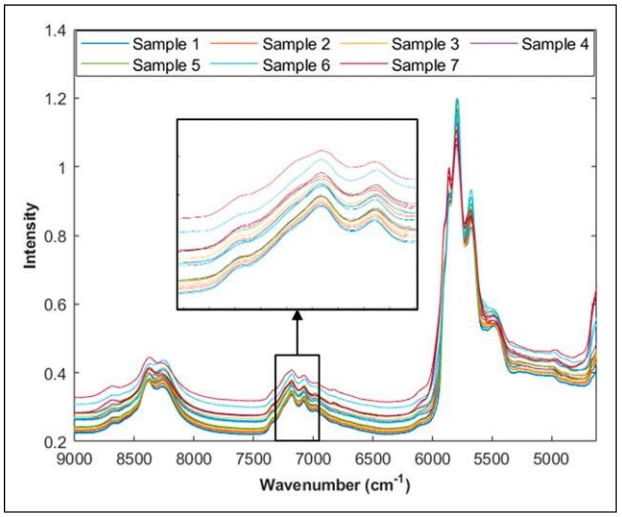



Figure 7. 21 raw spectra used in Case 2 evaluation over the entire spectral range. Legend: Color → samples analyzed. Black square → 21 raw spectra used in Case 2 evaluation magnified over the 7463 $cm^{-1}$ - 6993 $cm^{-1}$ spectral range.

method, which is the EMSC. In order to prevent loss of relevant information, the use of a minimum number of methods in the data preprocessing is generally recommended. Therefore, in this case, EMSC could be selected as the most efficient preprocessing method scenario to reduce the spectral variability due to the lack of reproducibility (see HCA dendrogram in Figure A2 in the supplementary section).

Figure A1 also shows that both the order of use and the combination of preprocessing methods can influence the final result. An example of this statement can be seen by comparing the scenarios from Case 1 without the measurement 35, where the NW-D method was combined with the MSC, PQN, SNV, and VSN methods. When the NW-D method (filtering category) is used before applying the other preprocessing methods (normalization category), the correct clustering is lower (about 30%) compared to when using the normalization preprocessing methods before the NW-D method. Comparing the same scenarios indicates that using complementary preprocessing methods does not always yield better correct clustering results than using a single method (NW-D+PQN → 38% Vs NW-D → 47%). This is an important consideration when using more than one preprocessing method.

The results obtained in this case show the proposed discrimination methodology's ability to evaluate the effectiveness of preprocessing methods in minimizing spectral variability in NIR measurements due to the lack of reproducibility. Moreover, it is worth mentioning that these results also illustrate the versatility of the proposed methodology for detecting potential anomalous data (outliers) caused by possible errors in spectrum acquisition.

As mentioned before, no detailed description for cases 2, 3, and 4 is presented; only their main results are shown. The detailed results of these cases are shown in the article's supplementary section.

## *Case 2—Repeatability*

Figure 7 shows the raw spectra used in the analysis of the variability caused by repeatability measurements. Analogous to case 1, the spectral variability generated by the physicochemical nature of the sample can be observed, presenting the same relationship (trend) with the properties reported in Table 1. However, the black square in Figure 7 shows that the variability caused by repeatability measurements seems to present a similar behavior to the variability generated by the combination of two different effects (additive and multiplicative), making the spectral differences of a single sample more evident in comparison with case 1.

In this second case, a total of 21 spectra (7 samples performed in triplicate) were analyzed to demonstrate the proposed methodology's ability to evaluate the effectiveness of preprocessing methods in minimizing unwanted

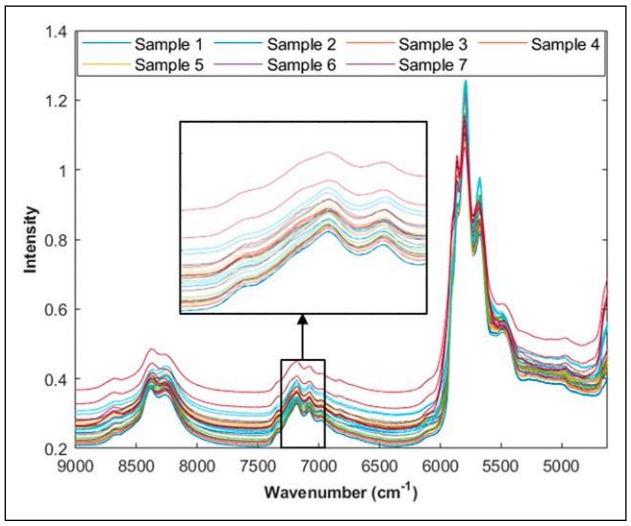



Figure 8. (a) 35 raw spectra used in Case 3 evaluation over the entire spectral range. Legend: Color → samples analyzed. Black square → 35 raw spectra used in Case 3 evaluation magnified over the 7463 $cm^{-1}$ - 6993 $cm^{-1}$ spectral range.

spectral variability due to the lack of repeatability. The correct grouping percentage achieved from all the preprocessing methods scenarios is shown in Figure A1 (orange square).

In the scenario that is assumed that the variability generated in the repeatability measurements do not have a significant impact on the grouping of sample measurements, that is, that no additional preprocessing methods are needed besides the data centering, only 48% of the measurements were correctly grouped (see Figure A1 (orange square) mean center scenario). Therefore, it can be preliminarily concluded that preprocessing methods to reduce spectral variations due to repeatability are needed. From the PCA analysis (data not shown) of this scenario (centered data), it could be determined that the first repetition of each sample differs significantly from its second and third repetitions. However, all of them were within the Q residual and Hotelling's $T^2$ test thresholds (data not shown). Therefore, they cannot be considered as anomalous data (outliers).

From all the evaluated scenarios, fourteen achieved a correct grouping of 100% (AWLS-B+MSC, AWLS-B+SNV, Dtd+MSC, Dtd+PQN, Dtd+SNV, EMSC+PQN, EMSC+SNV, MSC+AWLS-B, MSC+Dtd, PQN+EMSC, SNV+AWLS-B, SNV+Dtd, SNV+EMSC, and VSN). Among these scenarios, only one uses a single method, which is VSN (see HCA dendrogram in Figure A4). Therefore, VSN could be selected as the most effective preprocessing method to minimize the spectral variability due to the lack of repeatability. As in Case 1, Figure A1 shows that the order of use of the preprocessing methods affects the correct grouping result. Comparing the same scenarios as in Case 1, it can be reaffirmed that the results are more promising when the filtering methods are applied after the normalization methods.

From these results, it can be concluded that spectral variability due to repeatability has a lesser impact than those generated by reproducibility. Nevertheless, the proposed methodology demonstrated that the two cases' variability could be entirely compensated using an appropriate data preprocessing strategy.

## Case 3—Temperature effect

As previously mentioned in the introduction of the manuscript, sample temperature is one of the factors having significant impact on the NIR spectra acquisition. Figure 8 shows that the spectral variability generated by the sample temperature presents a behavior similar to the multiplicative effect. However due to the absorbance shift caused by the temperature increase, which prevents having a direct relationship between this parameter and the height of the acquired spectra, the spectral difference presents a non-linear growth.[45] This can be corroborated in the black square of Figure 8, where it is observed that with a temperature variation greater than 15°C, the spectral variability is more evident than when the delta in temperature is less than 15°C. This non-linear impact of the sample temperature on the spectrum acquisition could limit the performance of the different preprocessing methods evaluated.

In this third case, spectra acquired at 5 different sample temperatures (35 spectra) were analyzed to find the most effective preprocessing scenario to reduce undesired spectral variability due to the sample temperature. The correct grouping percentage achieved from all the evaluated preprocessing strategies is shown in Figure A1 (green triangle).

From Figure A1, it can be seen that if no other preprocessing method than data centering is applied, the percentage of correctly grouped measurements is 49%. This result reflects, as expected, the need to apply preprocessing methods to reduce variability caused by sample temperature. The clustering results shown in Figure A1 reveal that no evaluated preprocessing scenario could entirely compensate the spectral variability due to temperature variations, meaning that the entire accurate measurement grouping was not achieved in any scenario. The best performing scenario is the SG-D+SNV with an accurate grouping percentage of 80%. Although Figure A1 shows that the best performance scenario does not achieve the correct grouping of all 35 measurements, the results shown in Figure A5 (scenario SG-D+SNV) show that samples 1, 2, 3, 5, and 6 have an accurate grouping in all their measurements (5 out of 5 = 100%). The class affecting the overall measurement clustering is sample 4 (purple), which does not have any measurements grouped correctly. As in Case 1, it could be assumed that the whole misgrouping of sample 4 is due to the presence of some atypical data. However, no measurement was found above the Q residual and Hotelling's $T^2$ tests thresholds in any scenario (data not shown).

The results analyzed in this case show that sample temperature is a very influential parameter on the spectrum acquisition. Therefore, it is recommended to use a strategy that evaluates this variable's impact more thoroughly for a more efficient solution.[9,45]

## Case 4—cases 1, 2, and 3 combined

The parameters causing unwanted spectral variability evaluated in the 3 cases previously described are likely to occur simultaneously, mainly when online NIR measurement is used for real-time data analysis. For this reason, a fourth case was evaluated where the spectral variability generated by these three cases was combined.

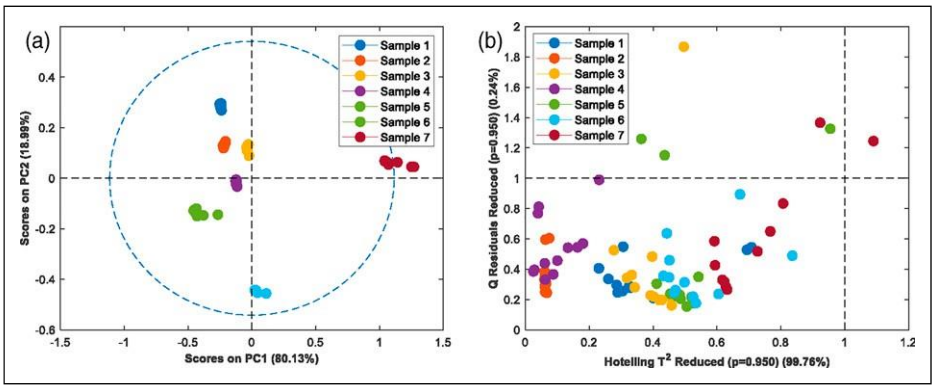



Figure 9. (a) Score plot of PC1 and PC2 for Case 4 with VSN+EMSC preprocessing and (b) reduced Qresidual and Hotelling's $T^2$ for Case 4 using 3 PCs with VSN+EMSC preprocessing. Legend: Color → samples measured (classes).

The dataset used in this case comprises 84 spectra in total. As in the 3 cases already studied, the Q residual and Hotelling's $T^2$ analyses were applied to the centered dataset to determine the possible presence of atypical data (data not shown). The measurement identified as an outlier in Case 1 (Sample 7, fifth acquisition day) once again exceeds the thresholds of the two tests. Thus, this measurement was removed, and the methodology proposed in this study was applied to the other 83 measurements.

In this case, 5 scenarios (EMSC, EMSC+PQN, PQN+EMSC, SNV+EMSC, and VSN+EMSC) had the best performance, where 80 out of 83 measurements were correctly grouped, that is, 96% of correct grouping. All 5 scenarios identified involve the use of EMSC, which, when evaluated individually, yields the same percentage of correct grouping (96%). It could be preliminarily inferred that for this case, the normalization methods do not give any additional improvement, and thus EMSC would be selected as the most effective preprocessing method scheme. Nevertheless, the use of the proposed methodology allows a more detailed analysis to determine at what level each type of variability is minimized in each of the 5 scenarios mentioned. In this way, the most effective preprocessing scheme's determination can be done more conveniently according to the researcher's objective.

Figure 9 shows the PCA score plot and the Q residual and Hotelling's $T^2$ results for the VSN+EMSC preprocessing method scenario. The three measurements that were not correctly grouped correspond to the measurement at 80°C of samples 3 (highest value of residual Q test) and 4, and the measurement at 75°C of sample 4. From Figure A1, it is also concluded that the EMSC and VSN methods have a better performance when grouping the samples measured at different temperatures in Case 4 (circle shape—purple color, combined cases) than in Case 3 (triangle shape—green color, sample temperature case).

Although the conclusion made previously that no method can entirely compensate the measurement variability caused by sample temperature is reaffirmed, the reduction of this variability was quite considerable in delta temperature lower than 20°C.

To sum up the results obtained by using the analysis methodology proposed, it could be concluded that for cases 1 and 2 (reproducibility and repeatability), the variability affecting the measurement clustering was fully compensated using a single preprocessing method. On the other hand, cases 3 and 4 (sample temperature and combination of cases) needed the combination of two methods, and still, the correct grouping of all the measurements was not achieved. The sample temperature has a high impact on the spectrum acquisition. Therefore, it is recommended to use a methodology that evaluates this variable more thoroughly for a more efficient solution.[9,45]

## Results comparison

In order to validate the consistency and reliability of the proposed approach, the analysis methodology results were compared with those obtained by Wilks' lambda criterion.

Table 3 summarizes each case's most relevant results achieved by both the proposed approach and Wilks' Lambda.

The results shown in Table 3 validate the approach proposed in this study. It can be seen that the results between the two methodologies are comparable, except for Case 1, including all measurements, when the data have been only mean-centered. In this case, Wilks' Lambda value is close to 1, while the value obtained by the proposed methodology is 0.46. The difference may be attributable to the presence of the outlier identified in Case 1 and how each approach handles this type of data. While the Wilks' lambda criterion assumes that there is no presence of outliers in the analyzed dataset, the methodology used in this study provides a preliminary analysis of the dataset for the identification and removal of possible anomalous data. This premise can be supported by observing that the two approaches' results are comparable when the identified outlier is removed from the dataset (see Table 3—Case 1 (Measurement removed)).

Moreover, the proposed methodology provides a more detailed discrimination analysis in comparison with Wilks' lambda criterion. As a way of example, both the proposed approach and the Wilks' lambda results obtained from the individual evaluation of the nine preprocessing methods in Case 1 were compared. From Table 4, it can be seen that Wilks' lambda criterion presents no significant differences in 4 preprocessing methods (SG-D, EMSC, MSC, and SNV), which could lead to the conclusion that the 4 methods have equal effectiveness in minimizing the unwanted spectral variability. On the contrary, the proposed methodology identifies that out of these 4 methods, two are equally effective in minimizing the unwanted spectral variability (MSC

Table 3. % Results comparison using Wilk's Lambda algorithm.

| Case | Preprocessing method | Methodology (% Grouped/100) | Wilk's Lambda |
|---|---|---|---|
| Case 1 (All measurements) | Mean center | 0.46 | 0.95 |
| | EMSC | 0.85 | 0.96 |
| Case 1 (measurement 35 removed) | Mean center | 0.53 | 0.68 |
| | EMSC | 1.00 | 0.99 |
| Case 2 | Mean center | 0.48 | 0.69 |
| | VSN | 1.00 | 0.99 |
| Case 3 | Mean center | 0.49 | 0.55 |
| | Sav–Gol + SNV | 0.82 | 0.95 |
| Case 4 | Mean center | 0.51 | 0.66 |
| | VSN+EMSC | 0.96 | 0.99 |

Table 4. Case 1 detailed results comparison.

| Preprocessing method | Explored methodology (% Grouped/100) | Wilk's Lambda |
|---|---|---|
| Mean center | 0.53 | 0.68 |
| AWLS-B | 0.65 | 0.88 |
| NW-D | 0.56 | 0.73 |
| SG-D | 0.65 | 0.98 |
| Dtd | 0.65 | 0.90 |
| EMSC | 1.00 | 0.99 |
| MSC | 0.82 | 0.99 |
| PQN | 0.68 | 0.87 |
| SNV | 0.82 | 0.99 |
| VSN | 0.94 | 0.85 |

and SNV), the least effective is the SG-D, and the most effective preprocessing method, which achieves the maximum compensation of the unwanted spectral variability, is the EMSC. Therefore, the methodology used in this work could help to select the preprocessing method in a more precise and reliable way. Finally, the proposed methodology enables identification of measurements and samples that have been properly or poorly discriminated, information that the Wilks' lambda does not provide as it is a global measurement.

## Conclusions

The results obtained in this study show the capacity of the analysis methodology used to assess the effectiveness of preprocessing methods in reducing the undesired spectral variability of nearinfrared spectroscopy measurements in a more thoughtfully and detailed manner than other approaches based on the dataset variance analysis such as the Wilks' lambda criterion.

In this study, an original strategy not previously reported in the literature was proposed to evaluate and determine the effectiveness of different preprocessing methods in minimizing the unwanted spectral variability due to parameters related to the continuous and repetitive NIR spectra acquisition such as repeatability, reproducibility, sample temperature, and the combination of these three parameters.

It is essential to stress the twofold benefit of using the proposed methodology. On the one hand, the detailed discrimination analysis provides a significant aid in determining the most effective data preprocessing scheme. On the other hand, the methodology provides a preliminary data analysis step for identifying and removing the potential anomalous data from the dataset, thus improving the reliability of the final results.

The results obtained using the proposed analysis methodology suggest that the variability caused by repeatability and reproducibility can be fully corrected when using the adequate preprocessing scheme; however, no preprocessing scenario could entirely compensate the unwanted spectral variability caused by the sample temperature. Similarly, the detailed discriminant analysis employed in this study showed that the EMSC preprocessing method presents interesting and promising results in all cases.

The preprocessing scheme's ultimate selection should be conducted in a careful manner considering the researcher's objective. The proposed methodology offers an analysis strategy that could help determine the most effective preprocessing scheme more reliably.

The conclusions reached in this work promote further optimization and automation of the proposed methodology to improve its implementation in large datasets.

The strategy proposed was shown to work for a case study including seven different hydrocarbon samples but can be generally applicable in any study involving spectroscopic information analysis

### Acknowledgments

The authors would like to thank IFP Energie Nouvelles for providing the hydrocarbon samples obtained in their HCK pilot plant reactors and the facilities for spectra acquisition and data analysis. Thanks also go to Axel One for providing the spectrometer for the NIR spectra acquisition.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Jhon Buendia Garcia https://orcid.org/0000-0002-7240-2362

## References

1. Fernández-Novales J, López MI, Sánchez MT, et al. A feasibility study on the use of a miniature fiber optic NIR spectrometer for the prediction of volumic mass and reducing sugars in white wine fermentations. *J Food Eng* 2008; 89: 3.
2. Saeys W, Nguyen Do Trong N, van Beers R, et al. Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: a review. *Postharvest Biol Technol* 2019; 158: 110981.
3. Blanco M and Peguero A. Analysis of pharmaceuticals by NIR spectroscopy without a reference method. *Trac Trends Anal Chem* 2010; 29: 10.
4. de Beer T, Burggraeve A, Fonteyne M, et al. Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. *Int J Pharmaceutics* 2011; 417: 32–47.
5. Balabin RM, Lomakina EI and Safieva RZ. Neural network (ANN) approach to biodiesel analysis: analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. *Fuel* 2011; 90(5): 2007–2015.
6. Zanier-Szydlowski N, Quignard A, Baco F, et al. Control of refining processes on mid-distillates by near infrared spectroscopy. *Oil Gas Sci Technol Rev IFP* 1999; 54: 463–472.
7. Wahl PR, Pucher I, Scheibelhofer O, et al. Continuous monitoring of API content, API distribution and crushing strength after tableting via near-infrared chemical imaging. *Int J Pharm* 2017; 518: 1–2.
8. de Oliveira RR, Pedroza RHP, Sousa AO, et al. Process modeling and control applied to real-time monitoring of distillation processes by near-infrared spectroscopy. *Anal Chim Acta* 2017; 985: 41–53.
9. Chauchard F, Roger JM and Bellon-Maurel V. Correction of the temperature effect on near infrared calibration—application to soluble solid content prediction. *J Near Infrared Spectrosc* 2004; 12: 199–205.
10. Igne B, Hossain MN, Drennen JK, et al. Robustness considerations and effects of moisture variations on near infrared method performance for solid dosage form assay. *J Near Infrared Spectrosc* 2014; 22(3): 179–188.
11. Betz JM, Brown PN and Roman MC. Accuracy, precision, and reliability of chemical measurements in natural products research. *Fitoterapia* 2011; 82(1): 44–52.
12. Olivieri AC and Faber NM. Validation and error. In: SD Brown (ed) *Comprehensive chemometrics. Chemical and biochemical data analysis*. Amsterdam: Elsevier, 2009, pp. 91–120.
13. ISO. *Measurement management systems–Requirements for measurement processes and measuring equipment*. UNE, 10012, 2003.
14. Wülfert F, Kok WT and Smilde AK. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Anal Chem* 1998; 70: 1761–1767.
15. Abe H, Iyo C and Kawano S. A study on the universality of a calibration with sample temperature compensation. *J Near Infrared Spectrosc* 2000; 8: 209–213.
16. Hansen WG, Wiedemann SCC, Snieder M, et al. Tolerance of near infrared calibrations to temperature variations; a practical evaluation. *J Near Infrared Spectrosc* 2000; 8: 8–10.
17. Al-Besharah JM, Akashah SA and Mumford CJ. The effect of temperature and pressure on the viscosities of crude oils and their mixtures. *Ind Eng Chem* 1989; 28(2): 213–221.
18. Luo P and Gu Y. Effects of asphaltene content on the heavy oil viscosity at different temperatures. *Fuel* 2007; 86: 7–8.
19. Payri R, Salvador FJ, Gimeno J, et al. The effect of temperature and pressure on thermodynamic properties of diesel and biodiesel fuels. *Fuel* 2011; 90(3): 1172–1180.
20. Rinnan Å, van den Berg F and Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. *Trac Trends Anal Chem* 2009; 28(10): 1201–1222.
21. Savitzky A and Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem* 1964; 36: 1627–1639.
22. Martens H and Stark E. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *J Pharm Biomed Anal* 1991; 9: 625–635.
23. Barnes RJ, Dhanoa MS and Lister SJ. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl Spectrosc* 1989; 43: 772–777.
24. Rabatel G, Marini F, Walczak B, et al. Variable sorting for normalization. *J Chemometr* 2020; 34(2): e3164.
25. Gholizadeh A, Borůvka L, Saberioon MM, et al. Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil Water Res* 2016; 10: 218–227.
26. Jiao Y, Li Z, Chen X, et al. Preprocessing methods for near-infrared spectrum calibration. *J Chemometrics* 2020; 34: e3306.
27. Brown CD, Vega-Montoto L and Wentzell PD. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Appl Spectrosc* 2000; 54(7): 1055–1068.
28. Gerretzen J, Szymańska E, Jansen JJ, et al. Simple and effective way for data preprocessing selection based on design of experiments. *Anal Chem* 2015; 87 (24): 12096–12103.
29. Devos O and Duponchel L. Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. *Chemom Intell Lab Syst* 2011; 107(1): 50–58.
30. Allegrini F and Olivieri AC. An integrated approach to the simultaneous selection of variables, mathematical preprocessing and calibration samples in partial least-squares multivariate calibration. *Talanta* 2013; 115: 755–760.
31. Luthria DL, Mukhopadhyay S, Lin LZ, et al. A comparison of analytical and data preprocessing methods for spectral fingerprinting. *Appl Spectrosc* 2011; 65(3): 250–259.
32. Wilks SS. *The large-sample distribution of the likehoold ratio for testing composite hypotheses*. Mathematical Statistics, 1962.
33. Rokach L and Maimon O. *Clustering methods*. Springer-Verlag.
34. Bricker M, Thakkar V and Petri J. Hydrocracking in petroleum processing. In: *Handbook of petroleum processing*. Routledge, 2014, pp. 1–35.
35. Speight JG. Hydrocracking. In: *The refinery of the future*. Elsevier, 2011, pp. 275–313.

1. ASTM D1218 - 12. "Standard test method for refractive index and refractive dispersion of hydrocarbon liquids", https://www.astm.org/Standards/D1218.htm
2. ASTM D2887 - 19ae1. "Standard test method for boiling range distribution of petroleum fractions by gas chromatography", https://www.astm.org/Standards/D2887.htm.
3. Norris KH and Williams PC. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. Influence of particle size. *Cereal Chem* 1984; 61(Mar): 158–165.
4. Martens H and Naes T. *Multivariate calibration.* Chichester: Wiley, 1989.
5. Dieterle F, Ross A, Schlotterbeck G, et al. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* 2006; 78(13):4281–4290.
6. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemom Intell Lab Syst* 1987; 2: 37–52.
7. Chen LS, Paul D, Prentice RL, et al. A regularized Hotelling's T2 test for pathway analysis in proteomic studies. *J Am Stat Assoc* 2011; 106(496): 1345–1360.
8. Sasirekha K and Baby P. Agglomerative hierarchical clustering algorithm- a review. *Int J Sci Res Publ* 2013; 33: 1–3.
9. Murtagh F and Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J Classif* 2014; 31(3): 274–295.
10. Roger JM, Chauchard F and Bellon-Maurel V. EPO–PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemom Intell Lab Syst* 2003; 66(2): 191–204.

## Appendix

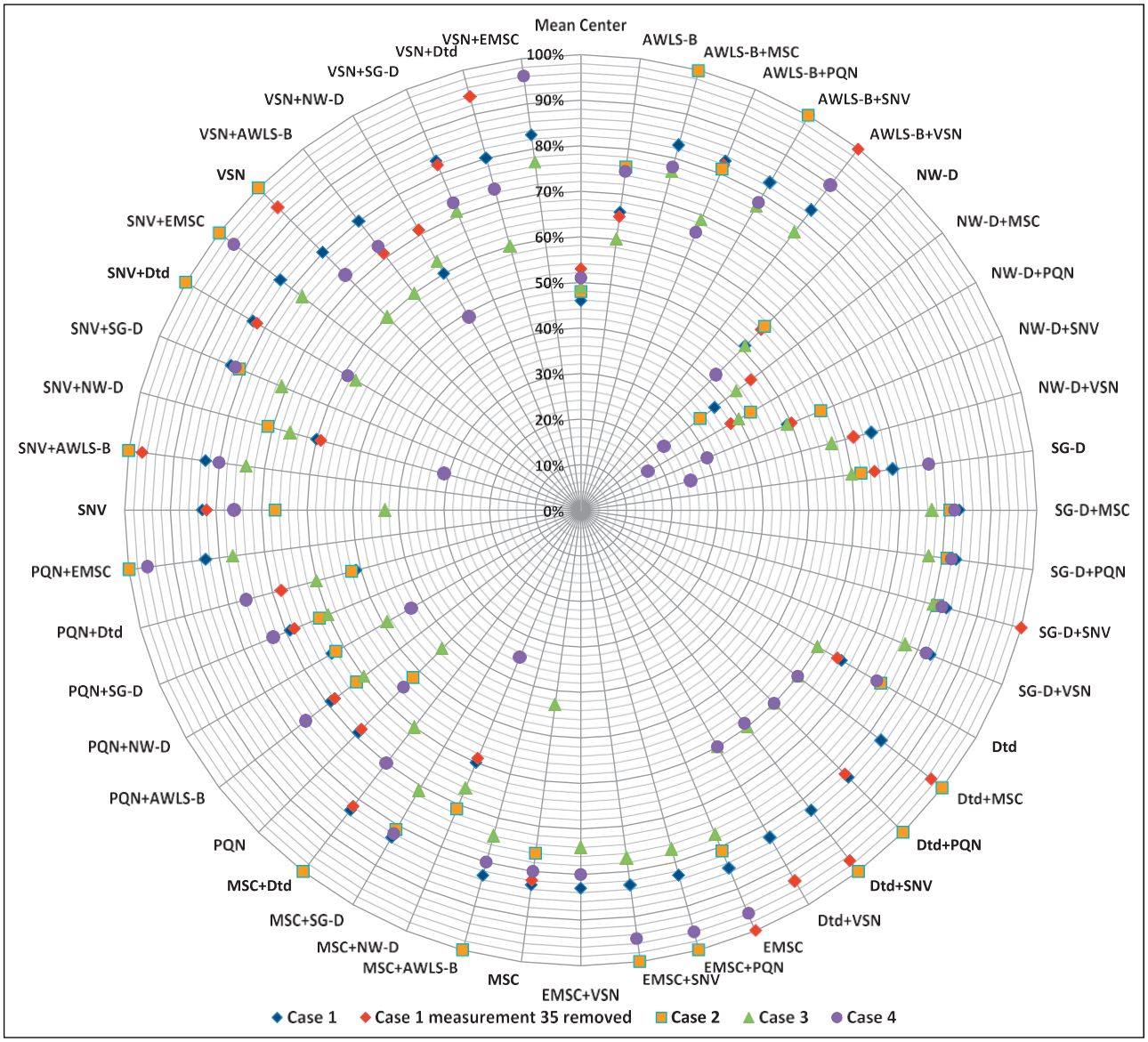
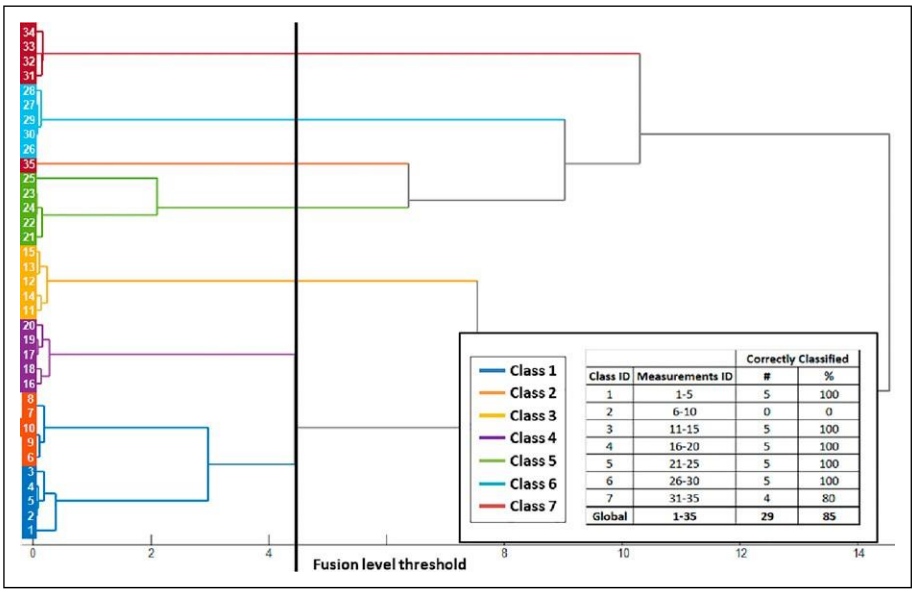
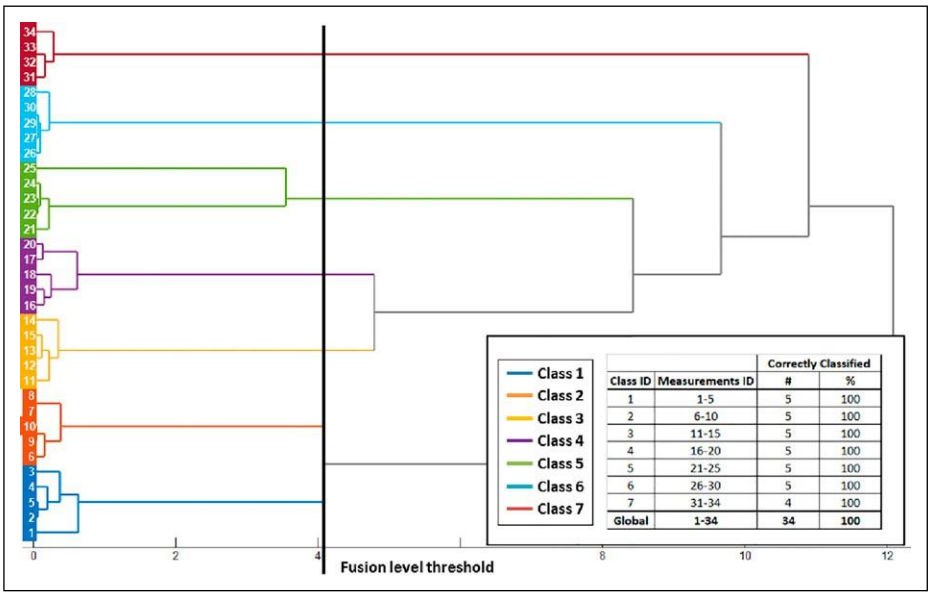



Figure A1. Comparative summary of the effectiveness of the preprocessing scenarios evaluated in the 4 case studies. Legend: Shapes and color → Case studies. Description: Circumferential lines → Percentage of correct grouping (0% Center—100% outer line). Radial lines → Evaluated preprocessing scenario, from left to right the order of use of the methods.



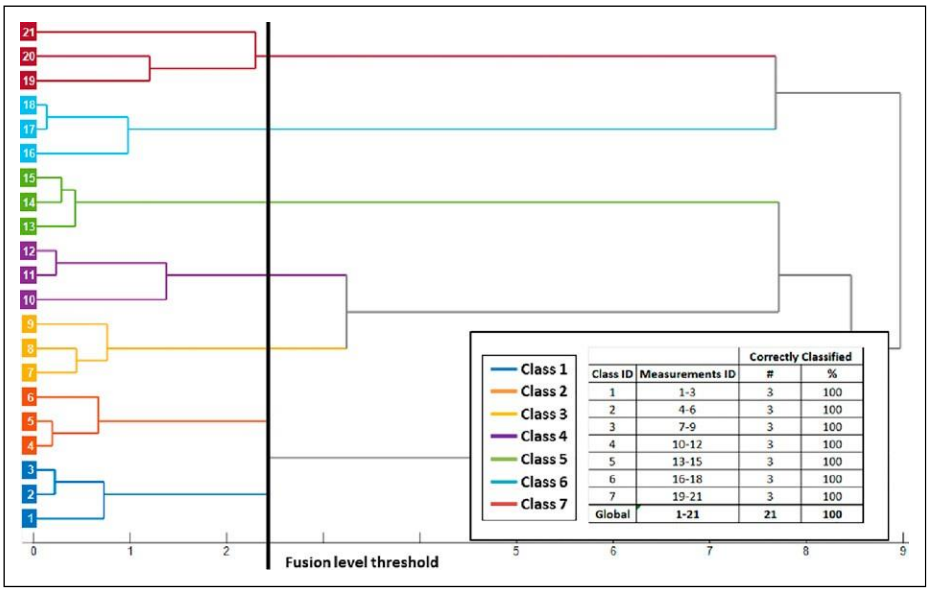

| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-5 | 5 | 100 |
| 2 | 6-10 | 0 | 0 |
| 3 | 11-15 | 5 | 100 |
| 4 | 16-20 | 5 | 100 |
| 5 | 21-25 | 5 | 100 |
| 6 | 26-30 | 5 | 100 |
| 7 | 31-35 | 4 | 80 |
| Global | 1-35 | 29 | 85 |

Figure A2. Dendrogram for Case 1 (all measurements) with EMSC preprocessing. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class



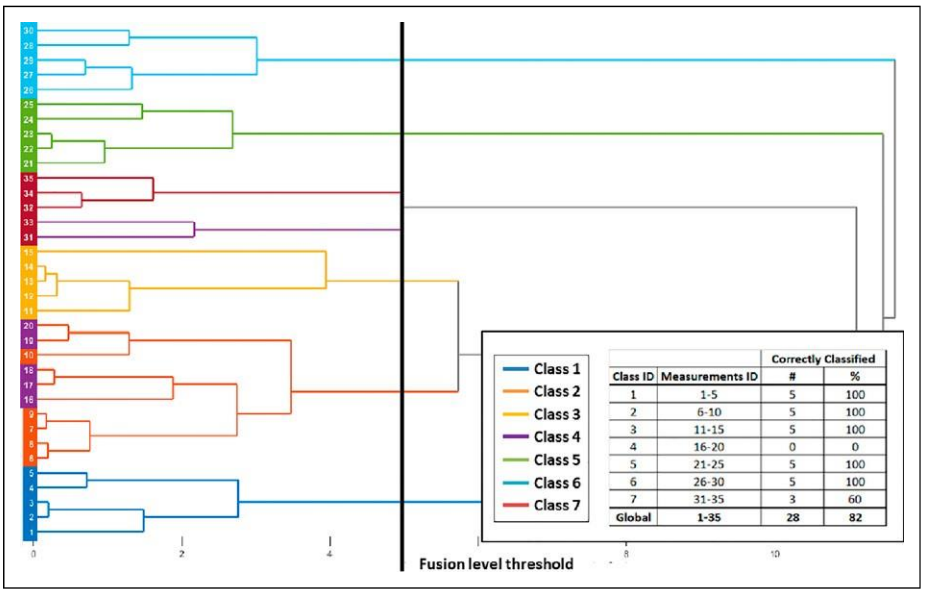

| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-5 | 5 | 100 |
| 2 | 6-10 | 5 | 100 |
| 3 | 11-15 | 5 | 100 |
| 4 | 16-20 | 5 | 100 |
| 5 | 21-25 | 5 | 100 |
| 6 | 26-30 | 5 | 100 |
| 7 | 31-34 | 4 | 100 |
| Global | 1-34 | 34 | 100 |

Figure A3. Dendrogram for Case 1 (removing measurement 35) with EMSC preprocessing. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class



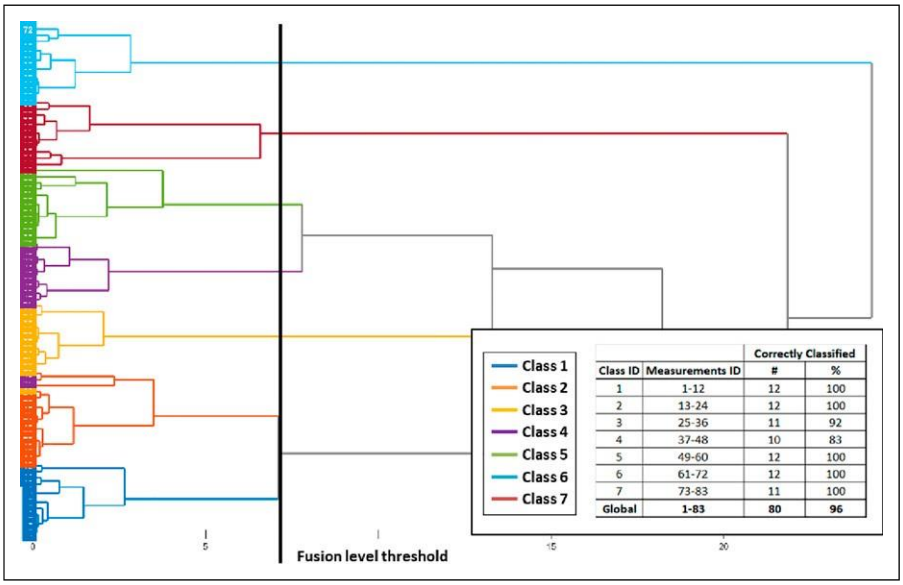

| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-3 | 3 | 100 |
| 2 | 4-6 | 3 | 100 |
| 3 | 7-9 | 3 | 100 |
| 4 | 10-12 | 3 | 100 |
| 5 | 13-15 | 3 | 100 |
| 6 | 16-18 | 3 | 100 |
| 7 | 19-21 | 3 | 100 |
| Global | 1-21 | 21 | 100 |

Figure A4. Dendrogram for Case 2 with VSN preprocessing. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class




| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-5 | 5 | 100 |
| 2 | 6-10 | 5 | 100 |
| 3 | 11-15 | 5 | 100 |
| 4 | 16-20 | 0 | 0 |
| 5 | 21-25 | 5 | 100 |
| 6 | 26-30 | 5 | 100 |
| 7 | 31-35 | 3 | 60 |
| Global | 1-35 | 28 | 82 |

Figure A5. Dendrogram for Case 3 with SG-D+SNV preprocessing. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class




| Class ID | Measurements ID | Correctly Classified # | Correctly Classified % |
|---|---|---|---|
| 1 | 1-12 | 12 | 100 |
| 2 | 13-24 | 12 | 100 |
| 3 | 25-36 | 11 | 92 |
| 4 | 37-48 | 10 | 83 |
| 5 | 49-60 | 12 | 100 |
| 6 | 61-72 | 12 | 100 |
| 7 | 73-83 | 11 | 100 |
| **Global** | **1-83** | **80** | **96** |

Figure A6. Dendrogram for Case 4 with VSN+EMSC preprocessing. Legend: Color samples measured (classes). Table: Correct grouping percentage for each class.

# Appendix 2: Publication #2. "Diesel cetane number estimation from NIR spectra of hydrocracking total effluent"

# Diesel cetane number estimation from NIR spectra of hydrocracking total effluent

J. Buendia Garcia[a,c], M. Lacoue-Negre[a,c], J. Gornay[a], S. Mas Garcia[b,c], R. Bendoula[b,c], J.M Roger[b,c]
[a] IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France
[b] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France

Corresponding Authors:
Marion Lacoue-Negre (marion.lacoue-negre@ifpen.fr)

## Abstract

The work shown in this paper offers a fast and efficient alternative for estimating the cetane number of the diesel obtained from the distillation of the hydrocracking total effluent. In this study, the estimation of this diesel property was achieved through a partial least squares regression (PLSR) model using only the NIR spectrum of the hydrocracking total effluent. For calibrating and validating the PLS model, it was used a database containing the NIR spectra acquired on 98 total effluent samples and the cetane number measured on the 98 diesel fractions recovered from each total effluent sample distillation. The database was divided into the calibration and test data sets using the Kennard-Stone algorithm. The regression model developed exhibited good performance in estimating the studied property with errors of calibration (1.3), cross-validation (2.2), and prediction (2.0), close to the reproducibility of the reference method (±3.6). The alternative method for diesel cetane number estimation discussed in this article evidences its feasibility in optimizing diesel fuel characterization by reducing the necessity of the total effluent distillation. Furthermore, the results also show the potential of the alternative proposed to be applied in predicting other properties of fuels obtained from the hydrocracking process.

## Keywords

Hydrocracking, Total effluent, Diesel, Cetane Number, Near-Infrared (NIR), Chemometrics.

# 1. Introduction

The shift in consumption from gasoline to diesel has led over the last 20 years to a strong worldwide increase in demand for middle distillates (kerosene and diesel) [1]. At the same time, the increasing heavy crude oil production [2] has resulted in low-quality feedstocks being processed. The outlined issues and the constant demand for high-quality products have raised the need for flexible refining processes that maximize the production of middle distillates from heavy feedstocks while ensuring their quality for compliance with environmental and commercial legislations[2,3]. Given its extensive flexibility in processing heavy feedstocks, the hydrocracking (HCK) process is essential in addressing the need described [4]. Moreover, as an extensively implemented process nowadays, it is the subject of ongoing research.

The research on the HCK process is conducted by implementing experimental designs in pilot plants and laboratory facilities under controlled conditions. The implemented experimentation contributes to determining the best process configuration by processing different types of residues, mostly vacuum gas oil

(VGO), under different operating conditions. In general, the experimentation is carried out in two main steps. In the first step, a hydrotreating stage (HDT) is applied to remove heteroatoms, saturate the olefins, and partially hydrogenate the aromatics. Subsequently, the hydrotreated effluent is sent to a reactor where, in the presence of a specific catalyst, the hydrocracking reactions occur [5] (See block #1 – Figure 1). In the second step, the liquid product obtained from the reaction section, known as total effluent, is distilled under atmospheric conditions to obtain the middle distillates, particularly diesel. These cuts are characterized using different standard norms such as the American Society for Testing and Materials (ASTM) and the International Organization for Standardization (ISO) (See block #2 – Figure 1). Finally, the analytical information obtained from this last step is gathered and analyzed to evaluate the impact of the operating conditions, including the catalytic system parameters, on the yield and quality of the diesel as a function of the processed feedstock.

In contrast to the reaction section, the characterization of the products is performed on a discontinuous time basis. Firstly, the laboratory analyses are conducted offline and are conditioned to the different laboratories' response times. Moreover, to perform the laboratory analyses based on the standards mentioned above, the physical product sample must be obtained from the total effluent distillation, which is also conducted in a non-continuous sequence. The products characterization is a fundamental task in the HCK process research. However, as previously discussed, the analytical workflow traditionally followed is both time- and volume-consuming. Therefore, a fast and efficient alternative for diesel fuel characterization is of great interest.

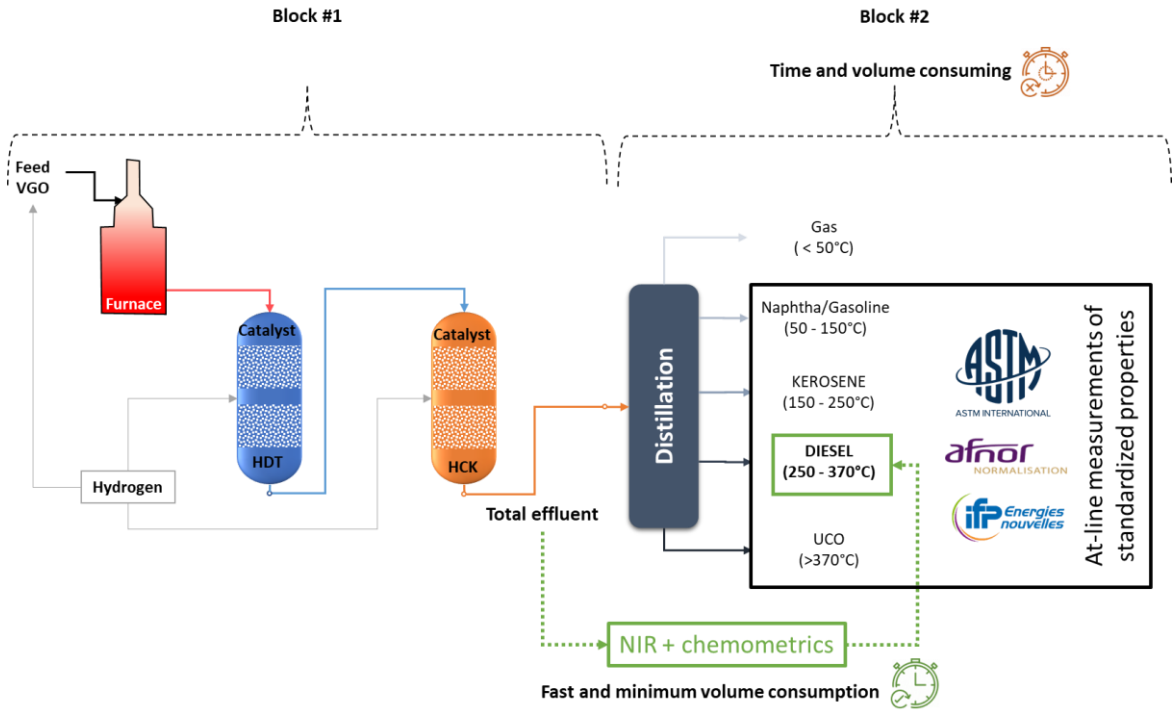



*Figure 1. Workflow scheme for the characterization of fuels obtained from the HCK process*

In the last decades, combining infrared spectroscopy analysis and chemometric methods has drastically increased for fuels characterization, from crude oils to refined cuts such as gasoline [6], diesel [7,8], biodiesel [7–10], and lubricants [11]. On the one hand, the main advantage of applying multivariate calibration methods to analytical techniques such as vibrational spectroscopy is both money- and time-saving. On the other hand, the sample volume required is quite low (up to a few milliliters) compared to some normalized methods generally used to characterize fuels. A recent review from Moro et al. [12] points out the growing use of infrared spectroscopy (IRS) to predict crude oil properties using chemometrics methods. To our knowledge, there is no existing equivalent review for other petroleum fractions. However, a plethora of interesting studies can be found showing the interest in using IRS and chemometrics to rapidly estimate fuel properties with statistical performance close to the reference methods [13–16].

Due to its extensive set of applications [17], NIR spectroscopy is particularly popular in laboratories to characterize fuels. Concerning diesel fuel, Hradecká et al. [15] recently demonstrated the feasibility of employing this vibrational technique to assess its quality. Using the partial least squares (PLS) algorithm, they estimated the kinematic viscosity, the cold filter plugging point, the pour point, and the sulfur and aromatics content from the NIR spectra acquired on different diesel samples. Each of the developed models enabled fast and reliable property predictions. Another recent study was developed by Yu et al. [18], where the estimation of diesel density from NIR spectra acquired on diesel samples was achieved using a "novel automatic model construction method." The resulting errors and squared correlation coefficients of the cited studies corroborated that an adequate application of chemometric methods on spectroscopic information leads to an accurate fuel properties estimation.

Among all the diesel fuel properties that can be investigated, the study shown in this article was focused on the diesel cetane number [19]. This property determines the ignitability of the diesel fuel using a standardized engine and a reference fuel. The cetane number is determined by comparing the ignition time of a mixture of cetane and hepta-methyl-nonane having the same ignition time delay as the tested sample. The cetane number on diesel is generally measured using the ASTM D613-01 standard [19], a destructive test requiring a significant sample volume (500 ml), and its response time is a couple of hours.

Regarding the diesel cetane number estimation using NIR spectroscopy, the most recent studies are reported by Zhan et al. [20] and Barra et al. [21]. In the first study, a least squares-support vector machine (LS-SVM) regression model was developed with errors of calibration (1.8) and prediction (2.0) lower than the reproducibility of the ASTM D613-01 standard method (~3.3). However, the squared correlation coefficients of calibration ($r^2c$) and prediction ($r^2p$) were quite low (0.66). In the second study, diesel cetane number estimations with prediction errors around 0.5 and an $r^2p$ value higher than 0.9 were achieved using a PLS regression model with 8 latent variables (LVs). Another study worth mentioning is the one developed by Zanier-Szydlowski et al. [22], who worked on predicting various fuel properties, including the diesel cetane number, developing a PLS model with a standard error of prediction (SEP) of 2.0.

All studies before-reported show that using NIR spectroscopy combined with proper chemometric methods in diesel properties estimation reduces the required sample volume and response time. However, the dependence on the distillation step of crude oil or HCK total effluent to obtain the diesel fraction and its subsequent characterization remains since the developed models are based on the NIR spectra acquired on the diesel cut. Therefore, aiming to go a step further in optimizing the analysis response time, this study presents an alternative for the cetane number estimation consisting of using the NIR spectra acquired on the HCK total effluent, avoiding the distillation step (see Figure 1). This main objective was achieved through four work steps. First, the total effluent samples obtained in different experimental tests of the HCK process conducted at a pilot level were identified and recovered. Next, the cetane number was measured on the diesel cuts corresponding to the total effluent samples. Then, the NIR spectra were acquired on the total effluent samples to finally perform all the necessary chemometric analysis, which included the preprocessing of the information and the calibration of the predictive model. To our knowledge, no comparative research has been reported.

# 2. Materials and methods

This section gives the origins and details of the sample physicochemical characterization. As a reminder, two sets of samples were considered: (i) the total effluents produced from HCK process reactors and (ii) the recovered diesel fractions.

## 2.1 Total effluent

In this study, 27 different feedstocks, mainly VGO, were processed in the HCK pilot plant units at IFPEN (Solaize, France) under various operating conditions involving different catalytic systems. The process variability ensured the physicochemical properties diversity of the 98 total effluent samples used in this research, as shown in Table 1. This table summarizes four relevant physicochemical properties of the obtained samples: the density,[23] the simulated initial boiling point (IBP), and distillation temperatures range to recover both 5% and 95% of sample distillate (Simulated Distillation T5 and T95)[24]. Table 1 also shows the fraction of the total effluent corresponding to the diesel cut.

*Table 1. Summary of physicochemical properties measured on the total effluent samples obtained from the hydrocracking process experimental tests.*

| | Méthod | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Density (g/mL)** | ASTM D1218-12[23] | 0.79 | 0.94 | 0.85 | 0.043 |
| **IBP (°C)** | ASTM D2887-19[24] | 38 | 205 | 106 | 42.1 |
| **SimDis T5 (°C)** | ASTM D2887-19 | 69 | 345 | 179 | 83.3 |
| **SimDis T95 (°C)** | ASTM D2887-19 | 401 | 585 | 503 | 44.6 |
| **Diesel yield (%)** | ASTM D2892-20[25] | 5.6 | 45.7 | 23.6 | 9.73 |

### Near-infrared analysis

Before NIR spectra acquisition, the samples were first heated in a water bath at 60°C in a closed flask for one hour and then manually shaken to ensure homogeneity. Subsequently, NIR analysis was performed on each of the total effluents obtained using a Falcata Lab6 immersion reflectance probe (Hellma GmbH & Co. KG, Müllheim – Germany) with an optical path fixed at 2 mm. A spectrometer NIRS XDS Process Analyzer (Metrohm, Villebon - France) recording wavelengths within the 800 - 2200 nm spectral range with a resolution of 0.5 nm was used to acquire the spectra. Each final spectrum obtained was the average of 32 scans performed on the sample. The software used with the spectrometer was VISION (Metrohm, Villebon - France).

## 2.2 Diesel

The diesel samples used in this study were recovered from the atmospheric distillation of each of the 98 total effluents according to the ASTM D2892-20[25] standard. The cetane number was measured on each diesel sample recovered using an IFPEN internal method, which estimates this property from diesel NIR spectra through a PLS model based on Zanier-Szydlowski et al. work [22], with a larger database and equivalent performance. The internal method outlined was developed using the cetane numbers measured using the ASTMD613-01 standard [19] analysis as the reference method and validated against the reproducibility limits defined by this norm. Table 2 summarizes the general statistical information of the cetane number, the density and the Simulated Distillation SimDis T5 and T95 of the diesel samples considered in this study.

*Table 2. General statistical information of the cetane number, density and simulated distillation measured on 98 diesel samples recovered from the total effluent distillation*

| | Method | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Cetane Number (CN)** | ASTM D5949 | 30.3 | 69.5 | 51.6 | 11.07 |
| **Density (g/mL)** | ASTM D1218-12 | 0.81 | 0.91 | 0.86 | 0.031 |
| **SimDis T5 (°C)** | ASTM D2887-19 | 213 | 258 | 245 | 9.1 |
| **SimDis T95 (°C)** | ASTM D2887-19 | 246 | 431 | 367 | 15.3 |

## 2.3 Modelling

An analysis to determine the best preprocessing scheme to be used was conducted. This study analyzed eight of the most common preprocessing methods applied to NIR spectra (see Table 3) [26] using an in-house MATLAB script. Each method was evaluated, taking their different parameter settings and possible combinations into account, based on the performance of different PLS regression models built using the root mean square error of cross-validation (RMSECV) and the squared coefficient of correlation ($r^2C$) as the figures of merit. For all models, the RMSECV was determined using the Venetian blind 10-fold.

*Table 3. Pre-processing method evaluated on the NIR spectra of the HCK total effluent*

| # | Category | Method | Acronym | Parameters |
|---|---|---|---|---|
| 1 | Normalization | Variable Sorting for Normalization [27] | VSN | Automatic calculation |
| 2 | | Standard Normal Variate [28] | SNV | |
| 3 | | Multiplicative Signal Correction [29] | MSC | Reference data = mean of data, whole spectral range |
| 4 | | Probabilistic Quotient Normalization [30] | PQN | |
| 5 | Filtering | Automatic Weighted Least Squares Baseline [26] | AWLS-B | |
| 6 | | Detrend [28] | Dt | Polynomial order (1-3) |
| 7 | | Extended Multiplicative Scatter/Signal Correction [31] | EMSC | Reference spectrum (basis to remove the scatter) = mean of each matrix generated, polynomial order = (1-4), whole spectral range, algorithm (CLS, ILS)* |
| 8 | | Savitsky-Golay Derivative [32] | SG-D | Window points (9-25), polynomial order = (1-4), derivative order (1-4) |

** CLS = Classical Least Squares, ILS = Inverse Least Squares.*

For building and testing the regression models, the database was split into two datasets using the Kennard-Stone (KS) algorithm[33]: the calibration set (70% samples), which was used in model calibration and internal validation (cross-validation), and the independent test set (30% samples), which was used in the performance evaluation of the final developed model. For each PLS model developed, the number of latent variables (LVs) with the lowest RMSECV was retained as long as the cross-validation and calibration error ratio (RMSECV/RMSEC) did not exceed 1.7. This criterion was established empirically through previous modelling results to avoid model overfitting. In addition, analogous statistics were calculated on the test set (RMSEP, r2P) to evaluate the model performance. The model errors were calculated using Eq. (1), where y_i and (y_i )̂ are the cetane number measured and predicted on sample i, respectively, and n is the number of samples. For the squared correlation coefficients calculation, Eq. (2) was utilized, where Cov and Var correspond to covariance and variance.

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

$$r^2 = \left(\frac{Cov(y,\hat{y})}{\sqrt{Var(y)Var(\hat{y})}}\right)^2 \quad (2)$$

The models were developed with the PLS_Toolbox V.8.9 (Eigenvector Research Inc. Wenatchee, WA, USA) and MATLAB V.2020b (The MathWorks, Inc., Natick, MA, USA).

# 3. Results and discussion

## 3.1 Preliminary spectral analysis

Before developing the regression models, a preliminary analysis of the NIR spectra was performed to determine the spectral range used. Based on the studies conducted by Yalvac et al.[34] and Kelly et al.[35], it was established that the spectral region between 1100 and 2200 nm provides the most informative spectral features for hydrocarbon samples. Figure 2a shows the absorbance spectra of the total effluent samples in this spectral range. Although assigning each band of a near-infrared spectrum to a hydrocarbon molecule is difficult, a global attribution can be done as follows: (A) the bands around 1200 nm correspond to the second overtone of the CH bands; (B) the bands in the spectral region 1300-1500 nm can be attributed to the combinations of vibrational modes for the stretching of CH bonds; (C) the bands in the spectral interval 1600-1850 nm correspond to the first overtone bands of -CH stretch in -CH2 and -CH3; (D) the bands around 2200 nm can be attributed to the combination absorption bands of -CH stretching bonds and C=C stretching bonds in the aromatic ring. According to the previously outlined information, it was decided to develop the models on the 1110-2200 nm spectral region.

The different preprocessing methods summarized in Table 3 were evaluated using the spectral range defined. The best performance scenario obtained for this study was the combination of the Standard Normal Variate (SNV) and the second derivative of Savitzky-Golay with a third polynomial order (SavGol[23,3,2]). The preprocessing scheme was completed by centering the matrix by columns (mean center). Figure 2b shows the spectra preprocessed where the four spectral zones identified before can be observed.

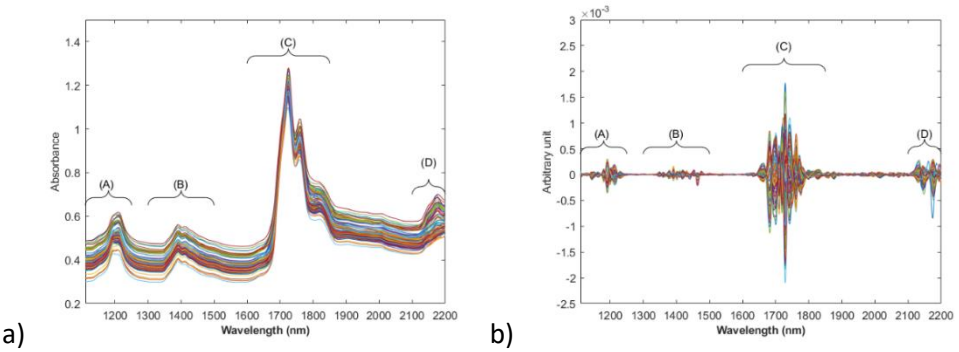



*Figure 2. a) NIR spectra in absorbance, b) NIR preprocessed spectra. Spectral range used in modelling (1110-2200 nm). Highlighted regions: (A)(1100-1250 nm), (B)(1300-1500 nm), (C)(1600-1850 nm), (D)(2100-2200 nm)*

## 3.2 Model performance analysis

After data preprocessing, a PLS model for the cetane number estimation was calibrated from the NIR spectra of 67 hydrocracked total effluent samples. The 67 corresponding diesel samples had a cetane number between 30.3 and 69.5. The external test set consisted of 31 total effluent spectra with an associated diesel cetane number ranging from 37.3 to 69.3. The score plot of the first two LVs of the developed PLS model shows a homogeneous distribution between the calibration and test samples (see Figure 3). This distribution ensures a representative evaluation of the model performance within the domain used in the model calibration. The distribution remains homogeneous throughout the other LVs (information not shown).

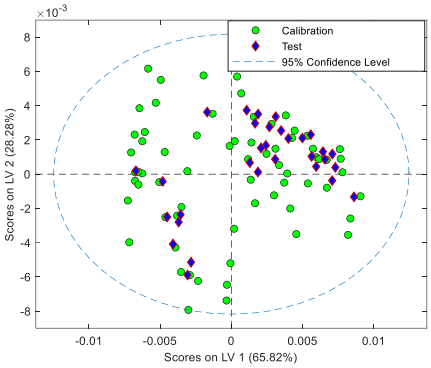



*Figure 3. Projection of the calibration and test sets over the first and second latent variables (Score-plot)*

The developed model uses 9 LVs to explain by about 99% the variance of the studied property. Considering the most recent studies regarding the estimation of diesel cetane number from NIR spectroscopy, the model developed in this study presents an RMSEP (2.0) comparable to the one obtained by Zhan et al.[20] (2.0) but presenting a better $r^2P$ (0.96 vs. 0.55). Additionally, compared to the regression method employed by them (LS-SVM), by using the PLS method in this study, the obtained model was interpretable, helping to understand the chemical information of the total effluent having an impact on the diesel cetane number. Regarding the study done by Barra et al. [21], which presents a lower RMSEP (0.42) using a PLS model of 8LVs, it should be noted that the data set used for testing their model is smaller (10 vs. 31) with a narrower cetane number range. The limited application range of the models reported in the two previously analyzed studies highlights another advantage of the model described in this article. While in the studies of Zhan and Barra the applicable model range is between 20.4-49.5 and 49-59, respectively, for the model developed is between 30.9-69.5.

Although the results of these studies are not rigorously comparable with the research shown in this paper due to the type of sample used for the NIR spectra acquisition (diesel vs. HCK total effluent), it can be observed that improvements in certain aspects are achieved. Furthermore, it is worth emphasizing that the alternative investigated in this study optimizes the diesel characterization response time, which was restricted by the distillation step. Finally, compared to ASTM D613-01 [19], the RMSEP of the developed model is below the reproducibility of all the cetane number ranges established by this standard.

In summary, using a PLS regression model with 9 LVs, it is possible to estimate the diesel cetane number from the spectroscopic information of the HCK total effluent with errors below the reproducibility limit of the IFPEN internal reference method (±3.6) and the ASTM D613-01 norm [19]. Moreover, the developed model ensures a reliable prediction throughout the entire range of property evaluation by presenting squared correlation coefficients higher than 0.95, showing a good correlation between the reference and predicted values. Table 4 shows the main information describing the chemometric model developed.

*Table 4 Statistical parameters and model information for predicting diesel cetane number (CN)*

| | |
|---|---|
| Regression method | PLS |
| Latent variables | 9 |
| X Explained Variance | 99.4% |
| Y Explained Variance | 98.6% |
| RMSEC | 1.3 |
| RMSECV | 2.2 |
| RMSEP | 2.0 |
| $r^2C$ | 0.986 |
| $r^2CV$ | 0.959 |
| $r^2P$ | 0.955 |
| Prediction Bias | -0.6 |

The satisfactory performance of the model obtained is reflected in the parity and residual plots shown in Figure 3a and Figure 3b, respectively. Figure 3a shows that out of the 31 samples used in the model test set, 30 were predicted between the lower and upper limits of the reference method reproducibility, resulting in a prediction effectiveness of approximately 97%. In turn, Figure 3b illustrates the homogeneous distribution of the residual values obtained in both the calibration and the test of the model, showing its homoscedasticity in the whole evaluation range of the studied property, and evidencing the absence of model overtraining.

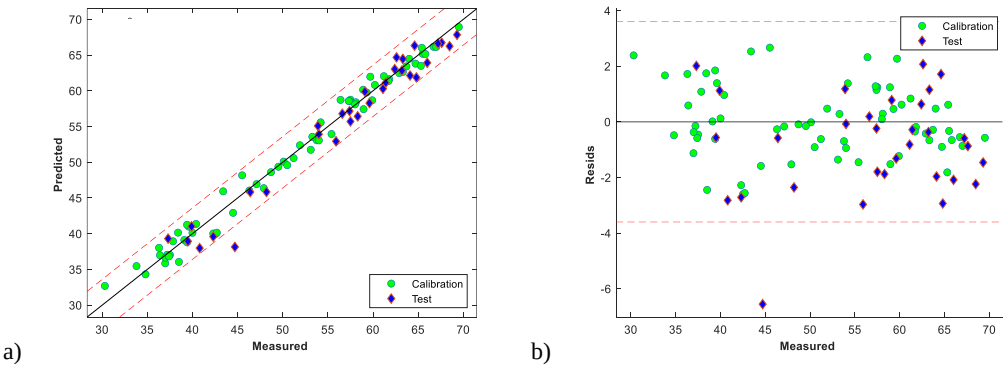



*Figure 3. a) Parity plot, b) prediction residuals plot of PLS model for predicting the diesel cetane number Red dotted lines: upper and lower limits of the reproducibility of the reference method (±3.6)*

A graphical analysis combining the Q residual and the Hotelling $T^2$ statistical analyses was performed to establish if the predicted sample outside the reproducibility limits of the reference method corresponds to an outlier. The Q residual test determines the samples with atypical behavior by measuring the difference between a sample and its projection into the LVs retained in the model [36]. If the residual Q value of a sample exceeds the unit, this sample can be considered a weak outlier, and its cause would be mainly related to the acquisition spectrum quality. Analogously, Hotelling's $T^2$ determines the atypicality of the samples using the measure of the variation in each sample within the model [36]. If the resulting test value of a sample exceeds the unit, it could be considered a strong outlier, and the cause would be mostly related either to the quality of the studied variable measurement or to the physicochemical properties of the sample. Finally, if a sample simultaneously exceeds the established thresholds of the two tests, the information from this sample

could substantially impact the model performance. Therefore, its use in the model should be reconsidered. Figure 5 shows the reduced Q residual and Hotelling $T^2$ analysis applied to the test set. Firstly, it can be observed in this figure that no sample is above the threshold of the two tests simultaneously. Secondly, two of the samples used in testing the model are above the threshold of the residual Q test. However, neither of these two samples corresponds to the sample predicted outside the limits. On the contrary, this sample is between the threshold limits of both tests (red point Figure 5). Consequently, it cannot be identified as an outlier. By a deeper analysis of this sample information regarding the operating and spectrum acquisition conditions, it was found that the total effluent sample analyzed was produced during a test with particular operating conditions in comparison to the rest of the sample set (feedstock with a high content of paraffinic carbon (>60%) processed under lower operating pressure). Thereby, the poor prediction of this sample could be attributed to the fact that the spectroscopic information used in the model calibration is not capturing the sample chemical description given by the particularity of the sample's origin. The present study focuses on estimating the studied property using NIR spectroscopy. The results indicate that this estimation is possible and that some external parameters can influence the prediction, such as operating conditions. This issue, related to the calibration robustness [37], could be addressed by developing predictive models that simultaneously use the information of the total effluent NIR spectra and the operating conditions employed in obtaining the sample.

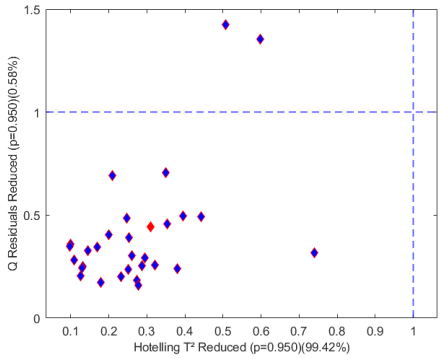



*Figure 5. Reduced Q residual and Hotelling $T^2$ analysis using a 9 LVs PLS model*

## 3.3 Model interpretation analysis

As mentioned before, one advantage of using the PLS regression method is to obtain predictive models helping to have a more detailed understanding of the effect that the different chemical compounds present in the sample may have on the estimation of the studied property. Figure 6 shows the PLS model loadings of the first 2 LVs, explaining 94% of the variance of the investigated property. This figure shows that the four zones previously identified influence the cetane number estimation. The zone between 1610 and 1810 nm is the one that presents the greatest impact. As mentioned formerly, this zone corresponds to the first overtone of the -CH stretching bands in -CH2 and -CH3. The behavior of the diesel cetane number is directly related to the type of isomerization, the length, and the amount of the identified linear hydrocarbons compounds. Therefore, the coherent relationship between the studied property and the chemical information extracted

from the NIR spectra acquired on the total effluent is demonstrated. This consistency suggests the possibility of applying the alternative proposed in this study to estimate other diesel properties.

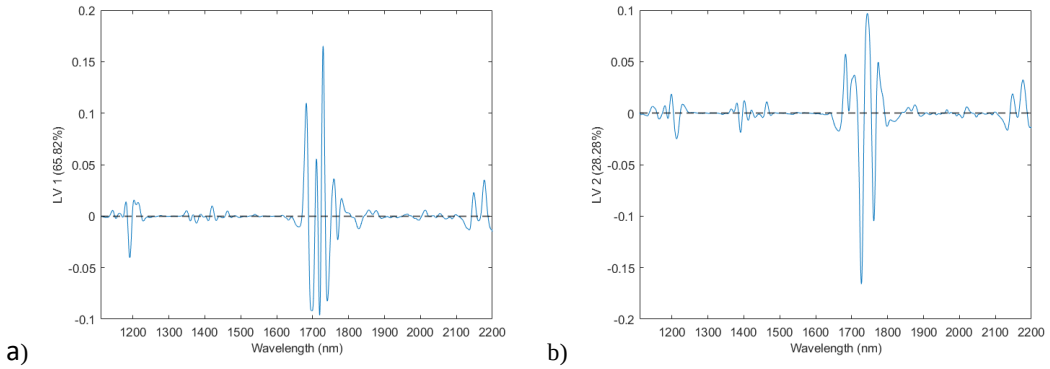



*Figure 6. PLS model loadings plot for a) 1st latent variable (65.8% of Y variable variance explained), b) 2nd latent variable (28.3% of Y variable variance explained)*

The previous description and results analysis validated the suitability of applying the alternative investigated in this article for estimating middle distillate properties with errors close to the reproducibility of the reference method. The diesel characterization alternative discussed in this study is based on exploiting the NIR spectra acquired on the HCK total effluent. The predictive model calibration represented a challenge during the research work due to the complex extraction and exploitation of the total effluent chemical information for correctly describing the studied property. Nonetheless, compared to the models for diesel cetane number estimation reported in the literature, the model developed in this study showed satisfactory performance. In addition, the model presents some further advantages concerning its homoscedasticity and its application range. Finally, as discussed in the introduction section, the interest in employing the total effluent NIR spectra was motivated by the need to go a step further in the response time optimization when characterizing the diesel fuel. Through the approach developed, this need is fully addressed as the distillation of the total effluent to recover the physical cuts is not required, offering the possibility of performing the properties estimation in real-time.

# Conclusions

The proper application of chemometric methods enables the physicochemical properties estimation of a crude oil cut using spectral information from another related product. This study developed a chemometric model for predicting the diesel cetane number using NIR spectroscopy information acquired on the total effluent obtained from the hydrocracking process. Hence, a fast and efficient alternative for fuel properties estimation was presented.

The PLS regression model obtained provides a reliable and fast estimation of the diesel cetane number with errors within the reproducibility of the reference method and correlation squared coefficients above 0.95. These results demonstrate the potential of the alternative investigated to minimize the required sample volume and the response time for property estimation by reducing the necessity to perform the total effluent

distillation. Furthermore, this optimization could lead to real-time and cost-effective research of the hydrocracking process by real-time estimating the studied property.

When estimating diesel properties using the spectroscopic information acquired on the total effluent, the predictive performance could be affected by the total effluent properties, which are impacted by parameters related to the feedstock quality and operating conditions. Therefore, it is important to address the model robustness constraint to ensure reliable performance over time and under different analytical conditions.

The study exposed in this paper highlights the wide application field of chemometrics, which facilitates the use of spectral information in the development of prediction models and enables the analysis and identification of atypical behaviors that fuel properties may have, helping to establish and understand the possible causes. Therefore, a better description of the influence of different process parameters and variables on the studied properties can be achieved, contributing to efficient process optimization.

The results obtained raise the prospect of using the alternative presented in this study for estimating other diesel properties as well as for properties prediction of different fuel products, namely, kerosene.

Finally, it should be highlighted that no regression model was found in the literature to predict diesel cetane number from NIR spectroscopy information of the hydrocracking total effluent, making this work the first one developed.

# Acknowledgments

The authors would like to thank IFP Energies Nouvelles for providing the total effluent samples from the HCK process reactors, the facilities for the distillation to obtain the diesel samples, and the facilities for spectra acquisition and data analysis. Thanks to Axel One Analysis for providing the probe used on the NIR spectra acquisition.

# Funding

This work was supported by IFPEN Energies Nouvelles

# CRediT authorship contribution statement

**J. Buendia Garcia**: Conceptualization, data curation, Writing - original draft. **M. Lacoue-Negre:** Conceptualization, Writing - original draft. **J. Gornay:** Conceptualization, Writing - original draft. **S. Mas Garcia:** Writing - original draft. **R. Bendoula:** Writing - original draft, **J.M Roger:** Conceptualization, Writing - original draft

# Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

# REFERENCES

[1] Fuels Europe. Statistical Report 2018. Study. Belgium; 2018.
[2] Marafi A, Albazzaz H, Rana MS. Hydroprocessing of heavy residual oil: Opportunities and challenges. Catalysis Today 2019;329:125–34. https://doi.org/10.1016/j.cattod.2018.10.067.
[3] Rana MS. Heavy Oil Refining Processes and Petrochemicals: A Role of Catalysis. Recent Adv Petrochem Sci 2017;2. https://doi.org/10.19080/RAPSCI.2017.01.555580.
[4] Elshout R, Bailey J, Brown L, Nick P. Upgrading the bottom of the barrel. Hydrocarbon Processing 2018, March 2018; Available from: https://www.hydrocarbonprocessing.com/magazine/2018/march-2018/special-focus-clean-fuels/upgrading-the-bottom-of-the-barrel.
[5] Vivas-Báez JC, Servia A, Pirngruber GD, Dubreuil A-C, Pérez-Martínez DJ. Insights in the phenomena involved in deactivation of industrial hydrocracking catalysts through an accelerated deactivation protocol. Fuel 2021;303:120681. https://doi.org/10.1016/j.fuel.2021.120681.
[6] Mabood F, Gilani SA, Albroumi M, Alameri S, Al Nabhani MM, Jabeen F. Detection and estimation of Super premium 95 gasoline adulteration with Premium 91 gasoline using new NIR spectroscopy combined with multivariate methods. Fuel 2017;197:388–96. https://doi.org/10.1016/j.fuel.2017.02.041.
[7] Balabin RM, Lomakina EI, Safieva R. Neural network (ANN) approach to biodiesel analysis: Analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. Fuel 2011, 2011:2007–15.
[8] Rocabruno-Valdés CI, Ramírez-Verduzco LF, Hernández JA. Artificial neural network models to predict density, dynamic viscosity, and cetane number of biodiesel. Fuel 2015;147:9–17. https://doi.org/10.1016/j.fuel.2015.01.024.
[9] Bemani A, Xiong Q, Baghban A, Habibzadeh S, Mohammadi AH, Doranehgard MH. Modelling of cetane number of biodiesel from fatty acid methyl ester (FAME) information using GA-, PSO-, and HGAPSO- LSSVM models. Renewable Energy 2020;150:924–34. https://doi.org/10.1016/j.renene.2019.12.086.
[10] Nabipour N, Daneshfar R, Rezvanjou O, Mohammadi-Khanaposhtani M, Baghban A, Xiong Q et al. Estimating biofuel density via a soft computing approach based on intermolecular interactions. Renewable Energy 2020;152:1086–98. https://doi.org/10.1016/j.renene.2020.01.140.
[11] Pinheiro CT, Rendall R, Quina MJ, Reis MS, Gando-Ferreira LM. Assessment and Prediction of Lubricant Oil Properties Using Infrared Spectroscopy and Advanced Predictive Analytics. Energy Fuels 2017;31(1):179–87. https://doi.org/10.1021/acs.energyfuels.6b01958.
[12] Moro MK, dos Santos FD, Folli GS, Romão W, Filgueiras PR. A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy. Fuel 2021;303. https://doi.org/10.1016/j.fuel.2021.121283.
[13] Morris RE, Hammond MH, Cramer JA, Johnson KJ, Giordano BC, Kramer KE et al. Rapid Fuel Quality Surveillance through Chemometric Modelling of Near-Infrared Spectra. Energy Fuels 2009;23(3):1610–8. https://doi.org/10.1021/ef800869t.
[14] Al Ibrahim E, Farooq A. Octane Prediction from Infrared Spectroscopic Data. Energy Fuels 2020;34(1):817–26. https://doi.org/10.1021/acs.energyfuels.9b02816.
[15] Hradecká I, Velvarská R, Dlasková Jaklová K, Vráblík A. Rapid determination of diesel fuel properties by near-infrared spectroscopy. Infrared Physics & Technology 2021. https://doi.org/10.1016/j.infrared.2021.103933.
[16] Feng F, Wu Q, Zeng L. Rapid analysis of diesel fuel properties by near infrared reflectance spectra. Spectrochim Acta A Mol Biomol Spectrosc 2015;149:271–8. https://doi.org/10.1016/j.saa.2015.04.095.
[17] Chung H. Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address. Applied Spectroscopy Reviews 2007;42(3):251–85. https://doi.org/10.1080/05704920701293778.
[18] Yu H, Wang X, Shen F, Long J, Du W. Novel automatic model construction method for the rapid characterization of petroleum properties from near-infrared spectroscopy. Fuel 2022;316. https://doi.org/10.1016/j.fuel.2021.123101.
[19] ASTM D613-01. Test Method for Cetane Number of Diesel Fuel Oil. West Conshohocken, PA: ASTM International; 2001. https://doi.org/10.1520/D0613-01.
[20] Zhan B, Yang J. Measurement of Diesel Cetane Number Using Near Infrared Spectra and Multivariate Calibration. Advances in Engineering 2017;100:270-247. https://doi.org/10.2991/icmeim-17.2017.41.

[21] Barra I, Kharbach M, Qannari EM, Hanafi M, Cherrah Y, Bouklouze A. Predicting cetane number in diesel fuels using FTIR spectroscopy and PLS regression. Vibrational Spectroscopy 2020;111:103157. https://doi.org/10.1016/j.vibspec.2020.103157.
[22] Zanier-Szydlowski N, Quignard A, Baco F, Biguerd H, Carpot L, Whal F. Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy. Oil & Gas Science and Technology - Rev. IFP 1999;54(4):463–72. https://doi.org/10.2516/ogst:1999040.
[23] ASTM D1218 - 12. Standard Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids; Available from: https://www.astm.org/Standards/D1218.htm.
[24] ASTM D2887 - 19ae1. Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography; Available from: https://www.astm.org/Standards/D2887.htm.
[25] ASTM D 2892-20. Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column). West Conshohocken, PA: ASTM International; 2020. https://doi.org/10.1520/D2892-20.
[26] Rinnan Å, van Berg F den, Engelsen SB. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends in Analytical Chemistry 2009;28(10):1201–22. https://doi.org/10.1016/j.trac.2009.07.007.
[27] Rabatel G, Marini F, Walczak B, Roger J-M. VSN: Variable sorting for normalization. Journal of Chemometrics 2020;34(2):205. https://doi.org/10.1002/cem.3164.
[28] Barnes RJ, Dhanoa MS, Lister SJ. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. Appl Spectrosc 1989;43(5):772–7. https://doi.org/10.1366/0003702894202201.
[29] Martens H, Naes T. Multivariate calibration. Chichester: Wiley; 1989.
[30] Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Analytical Chemistry 2006;78(13):4281–90. https://doi.org/10.1021/ac051632c.
[31] Harald M, Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. Journal of Pharmaceutical & Biomedical Analysis 1991;9.
[32] Abraham. Savitzky/M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 1964;36.
[33] Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics 1969;11(1):137–48. https://doi.org/10.1080/00401706.1969.10490666.
[34] Yalvac ED, Seasholtz MB, Crouch SR. Evaluation of Fourier Transform Near-Infrared for the Simultaneous Analysis of Light Alkene Mixtures. Appl. Spectrosc., AS 1997;51(9):1303–10. https://doi.org/10.1366/0003702971942303.
[35] Kelly JJ, Callis JB. Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. Anal. Chem. 1990;62(14):1444–51. https://doi.org/10.1021/ac00213a019.
[36] Le Mujica, Rodellar J, Fernández A, Güemes A. Q-statistic and T2-statistic PCA-based measures for damage assessment in structures. Structural Health Monitoring 2011;10(5):539–53. https://doi.org/10.1177/1475921710388972.
[37] Zeaiter M, Roger J-M, Bellon-Maurel V, Rutledge DN. Robustness of models developed by multivariate calibration. Part I. TrAC Trends in Analytical Chemistry 2004;23(2):157–70. https://doi.org/10.1016/S0165-9936(04)00307-3.

# Appendix 3: Publication #3. “NIR and 13C NMR data fusion to improve diesel cold flow properties prediction”

# NIR and 13C NMR data fusion to improve diesel cold flow properties prediction

J. Buendia-Garcia[a,c], S. Mas-Garcia[b,c], M. Lacoue-Negre[a,c], J. Gornay[a], R. Bendoula[b,c], J.M Roger[b,c]
[a] IFP Energies Nouvelles, Rond-Point de l'échangeur de Solaize, France
[b] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France

Corresponding Authors:
Marion Lacoue-Negre (marion.lacoue-negre@ifpen.fr)

## Abstract

This work presents a significant improvement in predicting diesel fuel properties by near-infrared (NIR) and $^{13}C$ nuclear magnetic resonance (NMR) data fusion modelling compared to separately generated models. In this study, the potential of three data fusion strategies (low-, mid-, and high-level) to improve the prediction of the diesel cold flow properties (pour point (PP), cloud point (CP), and cold filter plugging point (CFPP)) was investigated. The NIR and $^{13}C$ NMR spectra recorded on 84 total effluent samples obtained from hydrocracking process reactors were employed for developing the prediction models. For establishing the base case for comparison, partial least squares (PLS) regression models were developed using each data block separately. Then, data fusion models were built using the three strategies mentioned. The models were compared using the root mean square errors of calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP) as figures of merit. The mid-level data fusion modelling using the PLS scores as features extracted from each data block gave the best results. The RMSEP of the PP, CP, and CFPP was reduced by about 33%, 22%, and 20%, respectively, regarding the NIR model. The reduction in the RMSEC and RMSECV was between 19% and 43%. In addition, the squared correlation coefficients $r^2$ were also improved. All other data fusion strategies showed minor improvements. The results obtained illustrate the potential of the data fusion modelling to improve the diesel cold flow properties estimation.

## Keywords

Data fusion, Near-Infrared (NIR), Nuclear Magnetic Resonance (NMR), hydrocracking total effluent, diesel fuel, cold flow properties.

# 1. Introduction

Diesel characterization is fundamental for fuel assessment and refining process control and optimization [1–3]. Three of the most relevant properties measured in diesel fuel to determine its quality and performance are the pour point (PP) [4], the cloud point (CP) [5], and the cold filter plugging point (CFPP) [6]. These properties are also known as cold flow properties. In regions and countries where low temperatures (≤5°C)

are known to occur, the CP is the most considered parameter for the formulation of diesel fuel [7]. This property specifies the temperature at which the first paraffin or wax crystals appear. In turn, the PP seeks to determine the temperature at which the diesel stops flowing, and the CFPP establishes the temperature at which the crystallized wax begins to plug a standardized filter arrangement (simulating the fuel filter in a diesel engine) in such a way as to hinder the fuel flow.

The measurement of diesel cold flow properties is typically carried out using different laboratory standard norms such as the American Society of Testing Materials (ASTM) [4] and the International Organization for Standardization (ISO) [5, 6]. Due to the destructive and time-consuming nature of some of these standards, a common practice in the oil industry is the development of mathematical models as an alternative in estimating the properties using other analytical information from the analyzed stream [8]. In the last decades, an increasing interest in using vibrational spectroscopy to develop such models with statistical performance close to the reference methods has been reported [9].

The exploitation of information extracted from spectroscopic techniques for diesel characterization is feasible since these techniques can describe the behavior of hydrocarbon molecules present in a fuel sample. For instance, infrared spectroscopy (IR) provides information about the interactions of the C-H, C=C, aromatic =C-H, N-H, and O-H bonds. Some spectroscopic model developments can be found in the studies done by Hradecká et al., [10] where 7 diesel properties, including CFPP and PP, are predicted from the NIR spectra of diesel, and by Pasadakis et al., who developed regression models for diesel CP and PP prediction using MIR spectra acquired on diesel samples [11]. These studies show the complexity and difficulty of obtaining optimal regression models to predict diesel cold flow properties.

Although NIR spectroscopy provides information about the interactions of hydrocarbon molecules, this molecular information is quite general and may not be informative enough, yielding models with limited predictive performance. Therefore, combining the information from this technique with complementary and synergetic information such as the $^{13}C$ NMR spectroscopy, which, compared to NIR, contains more detailed information of the molecular interactions and bonds present in the sample, could help to improve the estimation of the fuel physicochemical properties [12–14].

Simultaneous use of information obtained from different analytical techniques for developing predictive models is generally referred to as data fusion. In their recent work, Azcarate et al. [15] detail the different strategies and applications of data fusion according to the structure of the data employed. This type of data manipulation generally employs different strategies known as fusion levels (low-, mid-, and high-level) [16]. The low-level fusion consists of using the information from the blocks directly in the development of the model either by simple concatenation of the blocks or using decomposition or factorization methods on one block regarding another [17]. At the mid-level fusion, a step of feature extraction from each dataset is

performed first through statistical analyses such as PCA and PLS for their later fusion by simple concatenation [18]. Finally, the high-level fusion combines the decisions or results obtained from developed prediction models separately with each data block [19].

Although data fusion methodology has been mostly used in the food industry [20], this type of modelling has also been employed in the oil industry. For example, Dearing et al. [13] employed the low-level data fusion strategy to estimate crude oil's API gravity and hydrogen content from three spectral signals (Raman, IR, NMR). They showed that compared to the individual models generated from each data block, data fusion reduced the prediction error by about 50%. Another study that employed data fusion for predicting the API gravity of crude oil was developed by Muhammad et al. [21]. In their study, they used two 1H NMR signals acquired at different relaxation times. The standard error of prediction (SEP) had a reduction of about 30% when the data fusion was performed. Recently Moro et al. [14] evaluated the three data fusion strategies described previously to predict seven crude oil properties (sulphur content (S), total nitrogen content (TN), basic nitrogen content (BN), total acid number (TAN), saturated (SAT), aromatic (ARO) and polar (POL) contents) using NIR, $^{1}H$ and $^{13}C$ NMR spectra acquired on crude oil samples. Their study showed that for all properties, the best data fusion strategy was the mid-level using the PLS scores.

Regarding using data fusion strategies for petroleum cuts characterization, Li et al. [22] compared the low-level and mid-level strategies to improve the estimation of methanol content in gasoline samples by fusing Raman and NIR spectral information. Showing improvements in the prediction error by around 50%, the mid-level strategy offered the best performance. Finally, Aguiar et al. [23] employed the mid-level data fusion strategy to improve the identification and characterization of adulterated diesel from NMR signals.

It is important to highlight that the regression models obtained in the studies mentioned above were developed using the spectroscopic information acquired on the stream being evaluated, for instance, predicting crude oil sulphur content using data fusion of MIR and NMR spectra acquired on the crude oil [14]. This approach is commonly used to develop prediction models; nevertheless, it may be useful for different applications to employ the approach of using spectroscopic information from one stream to predict the properties of another related stream [24]. To our knowledge, no data fusion models for predicting the diesel cold flow properties using the latter approach have been documented in the literature.

Considering the context previously outlined, the main objective of the work presented in this paper was to investigate the potential of the three data fusion strategies formerly described (low-, mid-, and high-level) to improve the prediction of the diesel cold flow properties using the NIR and the $^{13}C$ NMR spectra acquired on hydrocracking (HCK) total effluent samples. A detailed context on the modelling approach for predicting diesel properties from spectroscopic information of HCK total effluent can be found in [24].

# 2. Materials and methods

In this study, two sets of samples are considered: total effluents produced from HCK reactors (used for NIR and $^{13}C$ NMR spectra acquisition) and their distilled diesel cuts (used for cold flow properties laboratory analysis). This section gives the origins and details of the physicochemical characterization of the analyzed samples.

## 2.1 Total effluent

84 total effluent samples were obtained by processing 17 different vacuum gasoils (VGO's) in the hydrocracking (HCK) pilot plant units at IFPEN (Solaize, France) under different operating conditions and involving different catalytic systems. The process variability employed ensured samples having a wide diversity of their physicochemical properties, as shown in Table 1. This table also shows the fraction of the total effluent corresponding to the diesel cut.

*Table 1. Summary of physicochemical properties measured on the total effluent samples obtained from the hydrocracking process*

| | Method | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Density (gr/ml)** | ASTM D1218-12[21] | 0.7818 | 0.9209 | 0.8507 | 0.0369 |
| **IBP (°C)** | ASTM D2887-19[24] | 59.3 | 179.4 | 98.8 | 33.6 |
| **SimDis T5 (°C)** | ASTM D2887-19 | 80.9 | 335.5 | 165.8 | 69.0 |
| **SimDis T95 (°C)** | ASTM D2887-19 | 404.7 | 588.3 | 503.8 | 36.2 |
| **Diesel yield (%)** | ASTM D2892[111] | 6.9 | 45.2 | 25.1 | 6.3 |

### 2.1.1 Near-infrared analysis

To ensure the liquid state and homogeneity of the samples, they were heated in a water bath at 60°C in a closed flask for one hour and then manually shaken. Next, NIR analysis was performed on each of the total effluent samples obtained using a Fourier Transform Near-Infrared spectrometer (FT-NIR) MATRIX-F (Bruker, Optik GmbH, Ettligen - Germany), which with a resolution of 4 $cm^{-1}$ recorded 4148 wavenumbers within the range of 12000 - 4000 $cm^{-1}$ (833 - 2500 nm). Each final spectrum obtained was the average of 32 scans performed on the sample. An immersion Falcata Lab6 probe (Hellma GmbH & Co. KG, Müllheim – Germany) with an optical path fixed at 2 mm withstanding temperatures ranging from -40 °C to 200 °C was used to acquire the spectra. The software used with the spectrometer was OVP (OPUS Validation Program - Bruker, Optik GmbH, Ettligen - Germany).

### 2.1.2 $^{13}C$ NMR analysis

Prior to NMR analysis, the samples were heated at 70°C and manually shaken to ensure the homogeneity of the sample. 250µl of total effluent were mixed with 250µl of CDCl3 and 0,3mg of Fe(acac)3. $^{13}C$ NMR spectra were recorded at 50°C on a Bruker Avance 600 MHz spectrometer (Bruker Biospin Gmbh, Rheinstetten, Germany) operating at 150.9 MHz using a 5 mm QNP probe (time-domain 128k, 60° pulse, proton decoupling, acquisition time 56 min, relaxation delay 5 s, 512 scans). Zero filling and exponential line broadening (1 Hz) were applied before Fourier transform. The spectra were accurately phased and baseline adjusted. The $^{13}C$

NMR chemical shift of chloroform-d was set to 76.9 ppm as an internal standard.

## 2.2 Diesel

The cold flow properties were measured on the diesel recovered from the distillation of each of the 84 total effluents by using the ASTM D5949 method for the pour point (PP)[4], the ISO 3015 method for the cloud point (CP) [5], and the NF EN 116 method for the cold filter plugging point (CFPP)[6]. Table 2 summarizes the general statistical information of these three properties, as well as the density [25] and the simulated distillation temperatures range to obtain both 5% and 95% of sample distillate (Simulated Distillation T5 and T95) [26].

*Table 2. General statistical information of the cold flow properties, density and simulated distillation measured on 84 diesel samples obtained from the hydrocracking process*

| | Méthod | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Pour Point (PP)** | ASTM D5949 | -48 | -3 | -22.4 | 11.1 |
| **Cloud Point (CP)** | ISO 3015 | -44 | -1 | -18.2 | 7.99 |
| **Cold Filter Plugging Point (CFPP)** | NF EN 116 | -32 | 5 | -14.6 | 9.1 |
| **Density (gr/ml)** | ASTM D1218-12 | 0.8135 | 0.9106 | 0.8612 | 0.0275 |
| **SimDis T5 (°C)** | ASTM D2887-19 | 246.7 | 288.7 | 262.0 | 8.7 |
| **SimDis T95 (°C)** | ASTM D2887-19 | 344.0 | 408.3 | 361.1 | 9.1 |

## 2.3 Modelling

Before modelling, each data block was preprocessed to remove any noise that could affect the results. Table 3 summarizes the preprocessing schemes applied on the NIR block for each property studied.

*Table 3 Preprocessing methods scheme applied to the NIR block according to each cold flow property studied*

| Property | Methods | Parameters |
|---|---|---|
| **Pour Point (PP)** | Variable Sorting for Normalization (VSN)[120] | Tolerance = 0.0017, #Parameters = 3 |
| | Savitsky-Golay Derivative (SG-D)[118] | 19-point window, polynomial order = 3 First order derivative |
| **Cloud Point (CP)** | Standard Normal Variate (SNV)[162] | |
| | Extended Multiplicative Scatter/Signal Correction (EMSC) [163] | Reference spectrum (basis to remove the scatter) = mean of each matrix generated, polynomial order = 3, whole spectral range |
| **Cold Filter Plugging Point (CFPP)** | Standard Normal Variate (SNV) | |
| | Savitsky-Golay Derivative (SG-D) | 25-point window, polynomial order = 1 First order derivative |

Unlike the NIR spectra, the preprocessing applied to the NMR spectra was the same for the 3 diesel cold flow properties. First, the NMR spectra were aligned using the Interval Correlation Optimized (icoshift) algorithm [32]. Subsequently, these spectra were preprocessed using the Savitzky-Golay smoothing (15-point window, polynomial order = 0) and normalized (each variable is divided by the sum of the absolute values of all the variables for a given spectrum) [33].

For building and testing the models, the database was split into two datasets: the calibration set (70% samples), which was used both in model creation and cross-validation, and the independent test set (30% samples), which was used in the performance evaluation of the developed models. The database splitting was effected once only using the Kennard_Stone (KS) algorithm [34] applied on the NMR spectra, and the two resulting data sets were used in the models' development; in other words, the calibration and testing datasets are the same for all models.

In summary, PLS models for each property were developed from the spectroscopic information of each data block (NIR, $^{13}C$ NMR). These models were identified as individual models. Afterward, the spectroscopic information from each data block was combined using the low-, mid-, and high-level data fusion strategies. These models were identified as data fusion models. For each PLS model developed, the number of latent variables (LVs) with the lowest cross-validation error (RMSECV) was selected as long as the cross-validation and calibration error ratio (RMSECV/RMSEC) did not exceed 1.7. This criterion, established empirically through previous results, was used to avoid the overfitting of the models. For all models, the RMSECV was determined using the Venetian blind 10-fold.

From each developed model, the root mean square errors and the squared correlation coefficients of calibration and cross-validation (RMSEC, RMSECV, $r^2C$, $r^2CV$) were calculated to select the best performing model. In addition, analogous statistics were calculated on the test set (RMSEP, $r^2P$) to evaluate the performance of the models. A more detailed description of the modelling procedure is provided in the following sub-sections.

The models were developed with the PLS_Toolbox V.8.8 (Eigenvector Research Inc. Wenatchee, WA, USA) and MATLAB V.2019b (The MathWorks, Inc., Natick, MA, USA).

### Individual modelling

Following the preprocessing, splitting, and centering of the data sets, PLS regression models were calibrated for each property analyzed in this study using the NIR and the $^{13}C$ NMR blocks separately. These models, labeled as individual models, were established as the base case for comparison to investigate the potential of data fusion modelling.

### Data fusion modelling

As mentioned before, three data fusion strategies (levels) were used; a global description of the approach used in each of them is given below.

In low-level fusion, each data block was preprocessed separately (see section 2.3.1). Two approaches were used:

- Simple concatenation: the preprocessed data were scaled considering each block variance and subsequently concatenated according to rows into a single matrix. Finally, a PLS regression model was created using the fused matrix, determining the number of LVs by cross-validation.
- SO-PLS [35]: a PLS model was developed from one preprocessed data block. Next, the second data block was orthogonalized according to the scores of the already developed model. Finally, a model was obtained using the fused information from block 1 and orthogonalized block 2. It is essential to mention that in this second approach, the number of LVs in each block and the order of use of each of them (data block) can affect the result. For this reason, the evaluation of all possible combinations of latent variables of each block and their order of use was made to define the best prediction model (final model to be compared).

In mid-level fusion, a PCA was applied to each preprocessed block separately. The scores resulting from each PCA were concatenated into a single matrix, which was used to develop the final PLS regression model. To avoid arbitrariness in deciding the number of principal components (PCs) of each block to be fused, all possible combinations between 1 and 20 PCs for each block were evaluated when developing the prediction model. The best-performing model was selected based on the criteria previously mentioned for the final comparison. This model was identified as the "PCA mid-level fusion model". A second mid-level fusion model was developed following the same described procedure but using the PLS analysis scores conducted on each data block. This model was called the "PLS mid-level fusion model".

In high-level fusion, an individual PLS model from each data block was developed to predict each cold flow property. The values predicted were concatenated into a single matrix, which was used to develop a multiple linear regression (MLR) model. All possible combinations of latent variables (LVs) in developing the individual models were evaluated.

Although the final decision on the number of LVs and PCs to be used in the final model was made based on the performance of the models obtained, the maximum number of these parameters to be evaluated was limited to 20. In order to optimize the comparison of the different combinations of these parameters in the data fusion models, a MATLAB function was developed to perform this process automatically. At the end, 5 final data fusion models were selected, and their performance was compared to the individual models.

# 3. Results and discussion

The first step implemented in this study was defining the spectral range used. Buendia et al.[24] validated, based on the study of Yalvac [36] and Kelly et al. [37], that the region of the NIR spectrum that provides the most descriptive information of hydrocarbon molecule behavior is between 9000 and 4500 cm-1. This region has four main zones that describe (A) the second overtone of the CH bands (around 8300 cm-1), (B) the combinations of vibrational modes for the stretching of CH bonds (7600-6600 $cm^{-1}$), (C) the first overtone

bands of -CH stretch in -CH2 and -CH3 (6250-5400 $cm^{-1}$), and (D) the combination absorption bands of -CH stretching bonds and C=C stretching bonds in benzene ring (around 4500 $cm^{-1}$). Figure 1a shows these spectral regions identified over the whole spectral range of the raw spectra (12000-4000 $cm^{-1}$), while Figure 1b shows the raw spectra in the spectral range used for modelling properties.

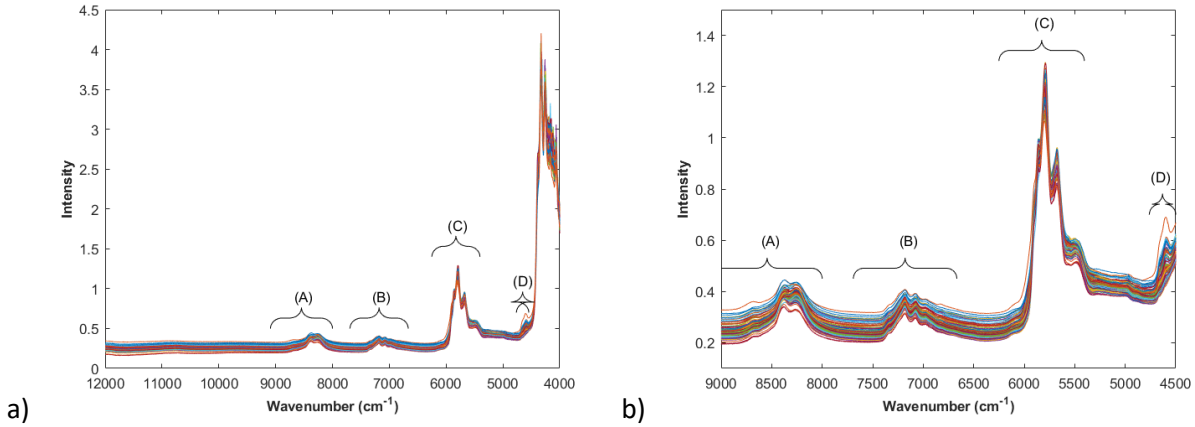



*Figure 1. a) NIR spectra full spectral range (12000-4000 cm-1), b) NIR spectra spectral range used in modelling (9000-4500 cm-1). Identified regions: (A)(9000-8000 cm-1), (B)(7600-6600 cm-1), (C)(6250-5400 cm-1, (D)(5800-4500 cm-1)*

Regarding the NMR analysis, in this study were used the chemical shifts in the region corresponding to the aliphatic carbons (0-60 ppm). The intensity of the peaks identified in the region corresponding to the aromatic carbons (100-150 ppm) does not significantly contribute to the developed models' performance. Figure 2 shows the NMR spectra of 3 total effluent samples in the 25-30 ppm chemical shift region. In this figure, a difference in the intensity of the peaks corresponding to the methyl branching in carbon β (22.8 ppm) and the methyl branching in a straight chain (29.9 ppm) [38] can be observed between the 3 samples spectra. This intensity pattern could provide complementary information related to the influence of this type of carbons on the diesel properties studied, contributing to the improvement of their estimation from spectral information.

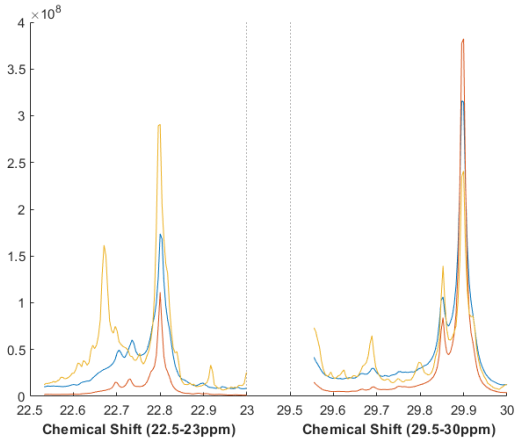



*Figure 2. NMR spectra peak intensity pattern. Left spectra: peaks in the 22-23 ppm chemical shit region. Right spectra: peaks in the 29-30 ppm chemical shit region*

Aft er defining the spectral range to be used in each data block, the best-performing models in each category

(individual and data fusion) were defined and compared to evaluate the potential of data fusion modelling to improve the diesel cold flow properties prediction.

## 3.1 Pour Point (PP) regression models

The results of the regression models developed for PP prediction are presented in Table 4. This table shows that, with an equal optimal number of LVs, the individual models (NIR, $^{13}$C NMR) present a similar calibration error (RMSEC) and equal prediction bias. However, the RMSECV is lower in the NIR model, while the RMESP is lower in the $^{13}$C NMR model. This error trend discrepancy shows that the stability of the individual models, which is determined by the consistency of the model errors, is limited for describing the diesel PP behavior.

*Table 4. Statistical parameters and model information based on individual and fused spectra for predicting diesel pour point (PP)*

| | | RMSEC | $r^2$C | RMSECV | $r^2$CV | RMSEP | $r^2$P | SEP | Bias Pre | LV* | LV/PC** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Individual | NIR | 4.1 | 0.854 | 5.5 | 0.736 | 5.7 | 0.754 | 5.7 | -0.3 | 5 | - |
| | $^{13}$C NMR | 4.0 | 0.864 | 6.0 | 0.696 | 4.9 | 0.821 | 4.9 | -0.3 | 5 | - |
| Low-level | Concatenation | 3.5 | 0.906 | 5.1 | 0.798 | 5.4 | 0.812 | 5.4 | -0.03 | 5 | - |
| | SO-PLS | 3.4 | 0.912 | 4.7 | 0.829 | 5.6 | 0.763 | 5.6 | **0.02** | - | 3 NIR,5 NMR |
| Mid-level | PCA | 4.2 | 0.866 | 5.0 | 0.810 | 5.7 | 0.746 | 5.7 | -0.2 | 5 | 15 NIR,13 NMR |
| | PLS | **2.6** | **0.940** | **3.3** | **0.907** | **3.9** | **0.899** | 3.9 | -0.3 | 4 | 2 NIR,14 NMR |
| High-level | - | 3.2 | 0.910 | 3.4 | 0.897 | 4.7 | 0.848 | 4.6 | -1.1 | - | 3 NIR,6 NMR |

**Latent variables used in the final model, ** Latent variables or principal components used in data fusion models*

Compared to the individual models, the best performing data fusion model is the mid-level model that uses PLS scores as features extracted from each block. This model is also known as PLS mid-level fusion model. Using this model, the RMSEC, RMSECV, and RMSEP are reduced by about 36%, 43%, and 33%, respectively. Another interesting finding is that the prediction bias is the same as the individual models. The consistency shown in these results demonstrates the potential of data fusion to better capture the descriptive information of each block of data, making the regression model more stable. The squared correlation coefficients $r^2$ were also improved.

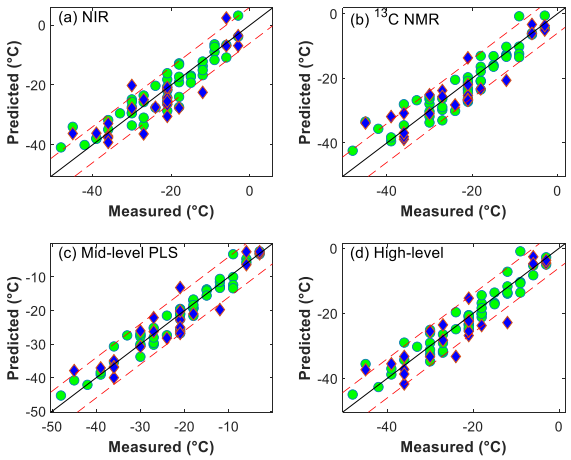



*Figure 3. Parity plots of the NIR, 13C NMR, PLS mid-level, and high-level fusion models for predicting the diesel pour point (PP). Series legend: circles  calibration samples, diamond  test samples. Red dotted lines: upper and lower limits of the reproducibility of the standard method (± 6°C) [4]*

The previous analysis can be corroborated in the parity plots shown in Figure 3, where the improvement in PP prediction when using data fusion is observed. Whereas the maximum percentage of predicted samples within the reproducibility limits of the reference method in the individual models is achieved in the NMR model with percentages of 84% in the calibration samples and 82% in the test samples, when using the PLS mid-level data fusion model these percentages increase to 100% and 88% for the calibration and test samples, respectively.

## 3.2 Cold Filter Plugging Point (CFPP) regression models

Table 5 summarizes the statistical parameters of the final models selected for diesel CFPP prediction. As for the PP prediction, the individual regression models for predicting diesel CFPP show similarity in calibration error and prediction bias. The lack of consistency in the errors is again evident as the same trend found in the PP estimation is observed, i.e., the RMSECV is lower in the NIR model, and the RMSEP is lower in the 13C NMR model. The major difference between the individual models is found in the squared coefficient of correlation $r^2$ of prediction, which is considerably lower in the NIR model, indicating the poor description of the CFPP behavior when using this data block. This low value is due to the inadequate prediction of the diesel samples having high CFPP values (> -2°C) (See Figure 4a).

*Table 5. Statistical parameters and model information based on individual and fused spectra for predicting diesel cold filter plugging point (CFPP)*

| | | RMSEC | $R^2C$ | RMSECV | $R^2CV$ | RMSEP | $R^2P$ | SEP | Bias Pre | LV* | LV/PC** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Individual | NIR | 4.6 | 0.743 | 5.3 | 0.664 | 6.1 | 0.513 | 5.9 | -1.4 | 5 | - |
| | $^{13}$C NMR | 3.8 | 0.828 | 5.8 | 0.619 | 4.9 | 0.863 | 4.8 | -1.2 | 6 | - |
| Low-level | Concatenation | 3.8 | 0.813 | 4.8 | 0.726 | 5.9 | 0.636 | 5.8 | -1.3 | 5 | - |
| | SO-PLS | 3.3 | 0.867 | 4.1 | 0.802 | 7.1 | 0.501 | 6.8 | -2.1 | - | 3 NIR,3 NMR |
| Mid-level | PCA | 4.1 | 0.799 | 4.8 | 0.726 | 6.0 | 0.628 | 5.8 | -1.5 | 5 | 14 NIR,9 NMR |
| | PLS | **3.4** | **0.865** | **3.5** | **0.814** | **4.4** | **0.874** | **4.3** | **-0.8** | 5 | 2 NIR,7 NMR |
| High-level | - | **3.4** | **0.866** | **3.4** | **0.859** | 5.4 | 0.675 | **5.3** | **-0.8** | - | 8 NIR,6 NMR |

**Latent variables used in the final model, ** Latent variables or principal components used in data fusion models*

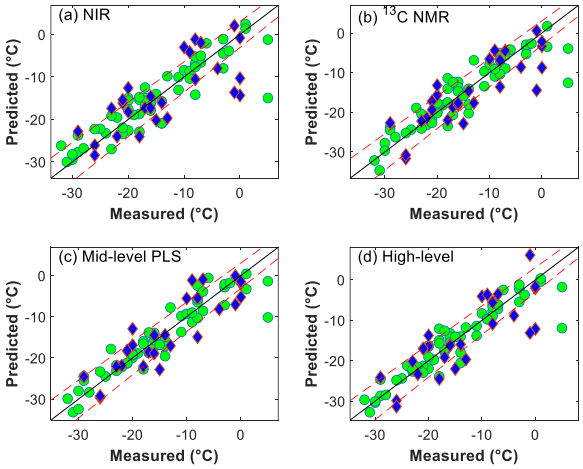



*Figure 4. Parity plots of the NIR, 13C NMR, PLS mid-level, and high-level fusion models for predicting the diesel cold filter plugging point (CFPP). Series legend: circles  calibration samples, diamond  test samples. Red dotted lines: upper and lower limits of the reproducibility of the standard method (±(3-0.06*CFPP)°C)[6]*

When using the PLS mid-level and high-level data fusion models, the RMSEC and RMSECV are reduced by an average of 19% and 37%, respectively. These models also reduce the prediction bias by 38%. However, the model with better stability and the better prediction error is the PLS mid-level fusion model. This model achieves a reduction in RMSEP of about 20% compared to the individual models.

By better capturing the relevant descriptive information of each block, the PLS mid-level fusion model reduces the influence that diesel samples with high CFPP values may have on the model performance. This model enhancement is observed in the parity plots shown in Figure 4, where the improved prediction of these samples (0°C > CFPP > -2°C) is evident. However, no evaluated approach allows to correctly model the samples with CFPP values > 0°C, whose predictions do not show a clear improvement. Since it was decided to keep the same calibration and test data set for the 3 cold flow properties of diesel, these values were not extracted in the present study. Figure 4 shows that out of the 84 diesel samples analyzed, only 2 have a high CFPP value (~5°C). One of the main reasons for a diesel sample to have this CFPP value is its high paraffin content. Since the estimation of this diesel property is being made from spectroscopic information acquired on the hydrocracking total effluent, and the percentage of samples having CFPP values greater than 0 is low (~2%), it is possible that the model developed does not capture the chemical information that these two samples can provide.

Although the cross-validation and prediction errors are significantly reduced, the percentage of predicted calibration samples within the reproducibility limits of the method is not significantly impacted. The single NMR model has 81% of predicted calibration samples within the reproducibility limits, while the PLS mid-level data fusion model has 82% of these samples within the limits. The percentage of predicted test samples within limits shows a slight improvement from 52% to 63%.

## 3.3 Cloud Point (CP) regression models

From Table 6, it can be observed that, unlike the two previously described cold flow properties, the errors of the regression models for the estimation of diesel CP present a better consistency and, therefore, better stability. Additionally, it can be observed from Table 6 that the prediction bias of the NIR model for CP estimation is the smallest compared to the individual models for estimating the cold flow properties of diesel. On the contrary, the prediction bias of the $^{13}C$ NMR model is significantly higher concerning all the models developed to estimate this property. This bias could be due to an insufficiently explained variance of the Y-block (85%) using 4 LVs, preventing an adequate description of the diesel CP behavior.

*Table 6. Statistical parameters and model information based on individual and fused spectra for predicting diesel cloud point (CP)*

| | | RMSEC | $R^2C$ | RMSECV | $R^2CV$ | RMSEP | $R^2P$ | SEP | Bias Pre | LV* | LV/PC** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Individual | NIR | 3.6 | 0.799 | 4.0 | 0.747 | 4.5 | 0.672 | 4.5 | **0.1** | 4 | - |
| | $^{13}$C NMR | 3.0 | 0.855 | 4.6 | 0.669 | 5.0 | 0.598 | 4.9 | -1.2 | 4 | - |
| Low-level | Concatenation | 2.6 | 0.916 | 4.0 | 0.792 | 3.3 | 0.824 | 3.3 | 0.4 | 6 | - |
| | SO-PLS | 2.4 | 0.908 | 3.4 | 0.816 | 4.3 | 0.764 | 4.3 | -0.2 | - | 7 NIR,1 NMR |
| Mid-level | PCA | 3.1 | 0.849 | 3.8 | 0.776 | 4.0 | 0.812 | 4 | -0.6 | 8 | 4 NIR,9 NMR |
| | PLS | **2.3** | **0.915** | **2.8** | **0.882** | **3.7** | **0.891** | 3.6 | -0.7 | 5 | 9 NIR,15 NMR |
| High-level | - | 2.7 | 0.887 | **2.8** | **0.875** | 3.8 | 0.816 | 3.8 | -0.2 | - | 6 NIR,6 NMR |

**Latent variables used in the final model, ** Latent variables or principal components used in data fusion models*

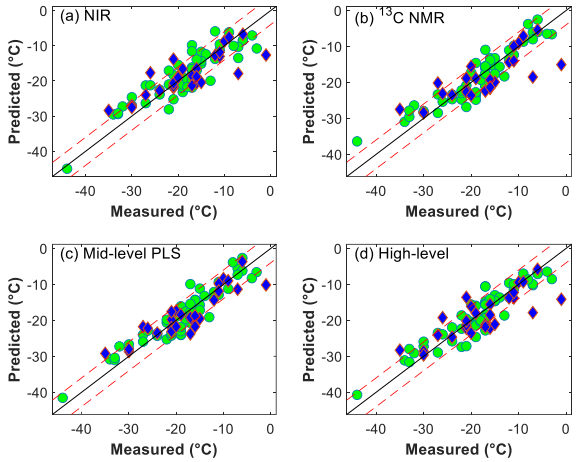



*Figure 5. Parity plots of the NIR, 13C NMR, PLS mid-level, and high-level fusion models for predicting the diesel cloud point (CP). Series legend: circles  calibration samples, diamond  test samples. Red dotted lines: upper and lower limits of the reproducibility of the standard method (±4°C)[5]*

Similar to the other two diesel cold flow properties, data fusion modelling, improves the prediction of diesel CP. The best performing models were the PLS mid-level and the high-level fusion models. These two models reduced the RMSECV and RMSEP by 34% and 22%, respectively. Although both models gave similar RMSEP, the squared correlation coefficient of prediction $r^2P$ is better in the PLS mid-level fusion model.

Once again, the potential of this model to optimally capture each data block relevant information for the prediction improvement of the studied properties is evident. Finally, Figure 5 shows the parity plots where the compensation of endpoints effect on the model performance can be observed again.

In comparison with the individual NMR model, the PLS mid-level data fusion model also improves the percentage of predicted samples within the reproducibility limits of the method, being more evident in the calibration samples (from 79% to 88%) than in the test samples (from 63% to 70%). It should be noted that the calibration and test data sets have been defined identically for all properties, and in the case of this property, the test basis may not be the best defined.

## 3.4 Final model comparison

A general model comparison was made to summarize the advantages of the data fusion modelling for predicting the diesel cold flow properties.

Compared to the individual models, the only property that showed a reduction in prediction error when using the low-level data fusion was the CP. For the other two properties, the prediction error reduction was negligible or even non-existent, i.e., the prediction error was higher in the low-level data fusion models than the individual models.

Regarding the mid-level data fusion strategy, no improvement in any cold flow property's calibration error was observed when using the PCA scores as features extracted from each data block; on the contrary, all calibration errors are higher than the individual models. Similarly, no reduction in PP and CFPP prediction errors was observed. The only variable that shows a reduction in this error was the CP. In contrast to the PCA mid-level fusion model, using the PLS regression scores as features extracted from each data block, significant reductions in all errors (RMSEC, RMSECV, RMSEP) were obtained for all properties studied. The $r^2$ correlation coefficients of all properties were also improved using the PLS mid-level fusion model. In Figure 4 and Figure 5, it can be observed that this mid-level model helps to reduce the effect that endpoints can have on the performance of the developed model.

The high-level data fusion modelling yielded lower calibration, cross-validation, and prediction errors than the individual models. Compared to the PLS mid-level model, the cross-validation error for all properties is fairly similar; however, the calibration error for the CP and PP properties and the prediction error for all 3 properties are higher in the high-level model. This is reflected in the squared coefficient of correlation, which is generally higher in the PLS mid-level fusion model.

In summary, the developed model comparison shows that the PLS mid-level and high-level models are the two best-performing models for diesel cold flow properties prediction. The results obtained from the high-level data fusion model are similar to those obtained with the PLS mid-level model; however, the latter presents greater stability in the errors, and it significantly reduces the negative impact that data endpoints can have on the model.

Figure 3 - 5 show the parity plots for the individual, mid-level PLS, and high-level models developed, where the upper and lower limits of the reference methods are shown (red dotted lines) (PP ±6°C[4], CP ±4°C[5], ±(3-0.06*CFPP)°C[6]).. These figures confirmed what was described in the previous paragraph, especially Figure 4 and Figure 5, where the improvement in upper endpoints prediction is observed, contributing to the prediction bias reduction, the $r^2$ coefficient increasing, and the model stability. This leads to the conclusion that the synergistic interaction between the two spectroscopic techniques (NIR and $^{13}C$ NMR) is better captured by the information extracted in the PLS mid-level fusion model than in the high-level model.

# Conclusions

This article investigates the potential of three data fusion strategies (low-, mid-, and high-level) to improve the prediction of three of the most important diesel fuel properties known as cold flow properties (pour point (PP), cloud point (CP), and cold filter plugging point (CFPP)). The data fusion modelling potential evaluation was conducted by comparing 7 final models: 2 individual models, 2 low-level data fusion models, 2 mid-level models, and 1 high-level model. The models were developed using 2 data blocks (NIR & $^{13}$C NMR).

The performance of the individual models (NIR and $^{13}$C NMR) is quite low in predicting diesel cold flow properties. The results of these models evidence the need to obtain more accurate predictions for these properties. Compared to these models, the low-level and mid-level data fusion using the PCA scores did not show significant improvements in the models' overall performance. Despite some punctual improvements in the errors and determination coefficients, there was a lack of stability and consistency in these models results. Based on the results analyzed, it can be determined that the use of these data fusion models, particularly the PCA mid-level model, is not desirable for predicting the diesel cold flow properties.

In contrast, the PLS mid-level and high-level data fusion models showed significant improvements in both errors and squared correlation coefficients compared to the individual models. Although the results obtained from the high-level and mid-level PLS models are comparable, the latter better captures the complementary information of each spectroscopic technique in the modelling. This is evident in the stability of the models, as well as in the reduction of the impact that data endpoints of the studied properties may have. Therefore, mid-level data fusion using PLS scores is the best strategy for developing the diesel cold flow properties predictive models.

The results demonstrated that the spectroscopic information from two different techniques could complement each other to improve the studied variables' behavior description. For the particular case of this study, it can be observed that the $^{13}$C NMR spectrum provides detailed and complementary information to the NIR spectra. Therefore, the data fusion of the two spectroscopic techniques employed in this study has potential use for fast and accurate properties prediction where errors of single models are higher than the reproducibility of the reference method.

In this study, models for predicting the diesel cold flow properties were developed using spectroscopic information acquired on the total effluent obtained from the hydrocracking process. The results obtained from both the individual and the data fusion models allowed validating the feasibility of using spectroscopic information from one stream (NIR and $^{13}$C NMR of the hydrocracking total effluent) to predict physicochemical properties of another related stream (cold flow properties of the diesel cut without distillation of the total effluent). No data fusion model was found in the literature to predict diesel cold flow properties from spectroscopic information of the total effluent obtained from the hydrocracking process,

making this work the first one developed.

# Acknowledgments

The authors would like to thank IFP Energies Nouvelles for providing the total effluent samples from the HCK process reactors, the facilities for the distillation to obtain the diesel samples, and the facilities for spectra acquisition and data analysis. Thanks also to Axel One Analysis for providing the spectrometer for the NIR measurements.

# CRediT authorship contribution statement

**J. Buendia-Garcia**: Conceptualization, Data curation, Writing - original draft. **J. Gornay:** Conceptualization, Writing - original draft. **M. Lacoue-Negre:** Conceptualization, Writing - original draft. **R. Bendoula:** Writing - original draft, **J.M Roger:** Conceptualization, Writing - original draft

# Declaration of conflicting interests

The Author(s) declare(s) that there is no conflict of interest

# REFERENCES

[1] A. Marafi, H. Albazzaz, M. S. Rana, Catalysis Today 2019, 329, 125–134.
[2] R. Mohan S, Recent Advances in Petrochemical Science 2017, 2, 19–22.
[3] M. A. Fahim, T. A. Alsahhaf, A. Elkilani in Fundamentals of Petroleum Refining, Elsevier, 2010.
[4] ASTM D5949, Standard Test Method for Pour Point of Petroleum Products (Automatic Pressure Pulsing Method), ASTM International, West Conshohocken, PA.
[5] NF EN ISO 3015 (2019-05-29), Petroleum and related products from natural or synthetic sources — Determination of cloud point.
[6] 2016-03-24 (NF EN 116), Diesel and domestic heating fuels - Determination of cold filter plugging point - Stepwise cooling bath method.
[7] C. Y. Tsang, V. S. Ker, R. D. Miranda, J. C. Wesch, Oil & Gas Journal 1988, 86.
[8] F. Al-Shanableh, A. Evcil, M. A. Savaş, Procedia Computer Science 2016, 102, 273–280.
[9] M. K. Moro, F. D. dos Santos, G. S. Folli, W. Romão, P. R. Filgueiras, Fuel 2021, 303, 121283.
[10] I. Hradecká, R. Velvarská, K. Dlasková Jaklová, A. Vráblík, Infrared Physics & Technology 2021, 103933.
[11] N. Pasadakis, S. Sourligas, C. Foteinopoulos, Fuel 2006, 85, 1131–1137.
[12] R. Legner, M. Voigt, A. Wirtz, A. Friesen, S. Haefner, M. Jaeger, Energy Fuels 2020, 34, 103–110.
[13] T. I. Dearing, W. J. Thompson, C. E. Rechsteiner, B. J. Marquardt, Appl Spectrosc 2011, 65, 181–186.
[14] M. K. Moro, Á. C. Neto, V. Lacerda, W. Romão, L. S. Chinelatto, E. V. Castro, P. R. Filgueiras, Fuel 2020, 263, 116721.
[15] S. M. Azcarate, R. Ríos-Reina, J. M. Amigo, H. C. Goicoechea, TrAC Trends in Analytical Chemistry 2021, 143, 116355.
[16] D. L. Hall, J. Llinas, Proc. IEEE 1997, 85, 6–23.
[17] Age K. Smilde, Iven Van Mechelen, Data Fusion Methodology and Applications 2019, 27–50.
[18] S. Agnieszka, E. Jasper, S. Ewa, B. Lutgarde, B. Lionel, Data Fusion Methodology and Applications 2019, 51–79.
[19] D. Ballabio, R. Todeschini, V. Consonni in Data Handling in Science and Technology, Vol. 31 (Ed.: M. Cocchi), Elsevier, Amsterdam, 2019.
[20] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Analytica Chimica Acta 2015, 891, 1–14.
[21] A. Muhammad, R. B. d. V. Azeredo, Fuel 2014, 130, 126–134.
[22] M. Li, J. Xue, Y. Du, T. Zhang, H. Li, Energy Fuels 2019, 33, 12286–12294.
[23] L. M. de. Aguiar, D. Galvan, E. Bona, L. A. Colnago, M. H. M. Killner, Talanta 2022, 236, 122838.
[24] J. Buendia Garcia, M. Lacoue-Negre, J. Gornay, S. Mas Garcia, R. Bendoula, J. M. Roger, Submitted to

Fuel 2022.
[25] ASTM D1218 - 12, Standard Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids, can be found under https://www.astm.org/Standards/D1218.htm.
[26] ASTM D2887 - 19ae1, Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography, can be found under https://www.astm.org/Standards/D2887.htm.
[27] ASTM D 2892-20, Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column), 2020, ASTM International, West Conshohocken, PA.
[28] G. Rabatel, F. Marini, B. Walczak, J.-M. Roger, Journal of Chemometrics 2020, 34, 205.
[29] S. Abraham, M. J. E. Golay, Analytical Chemistry 1964, 36.
[30] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Appl Spectrosc 1989, 43, 772–777.
[31] M. Harald, Edward Stark, Journal of Pharmaceutical & Biomedical Analysis 1991, 9.
[32] G. Tomasi, F. Savorani, S. B. Engelsen, Journal of chromatography. A 2011, 1218, 7832–7840.
[33] H. Martens, M. Høy, B. M. Wise, R. Bro, P. B. Brockhoff, Journal of Chemometrics 2003, 17, 153–165.
[34] R. W. Kennard, L. A. Stone, Technometrics 1969, 11, 137–148.
[35] T. Naes, O. Tomic, B.-H. Mevik, H. Martens, J. Chemometrics 2011, 25, 28–40.
[36] E. D. Yalvac, M. B. Seasholtz, S. R. Crouch, Appl. Spectrosc., AS 1997, 51, 1303–1310.
[37] J. J. Kelly, J. B. Callis, Anal. Chem. 1990, 62, 1444–1451.
[38] S. Verdier, J. A. Coutinho, A. M. Silva, O. F. Alkilde, J. A. Hansen, Fuel 2009, 88, 2199–2206.

# Appendix 4: Publication #4. “Variable selection and data fusion for diesel cetane number prediction”

# Variable selection and data fusion for diesel cetane number prediction

J. Buendia-Garcia[a,c], M. Lacoue-Negre[a,c], J. Gornay[a], S. Mas-Garcia[b,c],R. Bendoula[b,c], J.M Roger[b,c]
[a] IFP Energies Nouvelles, Rond-Point de l'échangeur de Solaize, France
[b] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France

Corresponding Authors:
Marion Lacoue-Negre (marion.lacoue-negre@ifpen.fr)

## Abstract

This study evaluates the potential of variable selection to improve the performance of data fusion between NIR spectroscopy information acquired on total effluent samples obtained from the hydrocracking process and their operating variables to estimate diesel cetane number. The evaluation conducted in this research was divided into four steps. First, predictive models were developed using each data block separately. Next, seven variable selection methods were applied on the NIR block, and eleven methods were applied on the process variable block. Then, with each data set generated from the variable selection analysis, single prediction models were generated and compared with those developed in the first step. Finally, data fusion was performed once the best variable selection method was defined for each data block. Two data fusion models were generated, a first using all the variables in the two blocks and a second using only the previously selected variables. In addition, the potential of the sequential and orthogonalized covariance selection (SO-CovSel) method was also analyzed. The results showed that the data fusion modelling using all variables from each data block improves the estimation of the diesel cetane number compared to single models (about 20% reduction of the RMSEP). However, using variable selection analysis before data fusion significantly improves the estimation of this property and leads to greater model stability regarding the RMSE's and $r^2$'s (about 47% of the RMSEP). The Covariance Selection (CovSel) method was the most efficient in the NIR data block, while for the process variable data block, it was the sequential backward floating feature selection method (SBFFS) that gave the best performance. The advantages offered by the variable selection resulted in having a more accurate prediction of the property and improving the analysis and understanding of the process by determining the variables that significantly impact the property studied.

## Keywords

Variable selection, Near-Infrared (NIR), process variables, data fusion, hydrocracking, diesel fuel, cetane number.

# 1. Introduction

The increasing use of analytical techniques such as vibrational spectroscopy for the rapid estimation of petroleum properties and their cuts has enhanced to some degree the analysis of refining processes providing valuable insights in their investigation and contributing to identify improvement and optimization opportunities [1,2]. Moreover, spectral modelling has been used for process monitoring and real-time decision-making based on the estimation of product properties. Some examples can be found in biodiesel

production reaction monitoring [3], gasoline property estimation to optimize the blending process [4], gasoline-ethanol blend distillation process control [5], and advanced crude oil refining planning [6].

Another source of data used for property estimation and process monitoring is related to the physicochemical information of the streams involved in the process and its operating conditions [7]. Generally, the quality of the streams is obtained by using various standardized methods, while different process sensors collect the information concerning the process operation. The estimation of product properties used in the analysis and understanding of the process must be done reliably and accurately, allowing to consider the different changes that may occur in the process operation. In the oil refining industry, the process variables most frequently used in the analysis are pressure, flow rate, temperature, and the characteristics of the catalytic systems used in the conversion processes [2].

The extensive availability of information and the increasing methodologies development observed in the last decades for exploiting data from different sources (analytical techniques and process variables) simultaneously, have helped to address the continuous need to improve the accuracy and stability of predictive models in several research and application fields, including the oil & gas segment. The model performance improvement using this type of data exploitation, known as data fusion, relies mainly on how is performed the interaction between the information obtained from the different sources. The simplest form of interaction is generally referred to as low-level fusion and involves the direct concatenation of the information contained in each data block. Two additional fusion levels such as mid-level, which consists of concatenating the features extracted from a statistical analysis performed on the different data blocks, or high-level, which uses the decisions or results obtained from prediction models calibrated on each data block, can be used for the data blocks interaction [8].

The application of data fusion has improved the properties estimation of both crude oil [9] and its cuts, particularly diesel [10,11]. However, studies in this area have mainly used information from spectroscopic techniques, such as infrared spectroscopy (IRS) and nuclear magnetic resonance (NMR). While the improvement in property estimation is evident, the performance of the models may be affected by the inherent variability of the processes. To overcome this drawback, the fusion of data related to the process operation and spectroscopic techniques is an alternative that is being used increasingly nowadays. Among the most recent works, de Oliveira et al. [12] using the mid-level data fusion strategy, evaluated the potential of using data simultaneously from process sensors and NIR spectra acquired on the product for process monitoring and control in three case studies ("Fluidized bed drying of pharmaceutical granules"," Polyester production process", and "Automated benchtop batch gasoline distillation") . In addition, Strani et al. [13] used data fusion between two sets of NIR spectra acquired on process streams and process operating variables to monitor polymer production. In these studies, the simultaneous and synergic use of heterogeneous data improved the monitoring and control of the analyzed processes.

The advantages of using process and analytical information simultaneously are straightforward. Nonetheless, using too many variables is risky and can lead to deteriorate the model performances, especially if a few samples are available. Some attention must be paid in this case, and a solution is to remove the information which does not provide an adequate description of the analyzed property behavior. Therefore, the determination of the most relevant descriptors is an important task in the optimization of the process analysis. This task, known as variable selection, can be implemented either for cost reduction in data acquisition (design and selection of sensors), for optimization of machine consumption, or for model optimization (accuracy or stability). An intrinsic advantage of performing variable selection is a better understanding of the interaction between the independent variables and the estimated property, leading to a better understanding of the evaluated process and its potential optimization.

Various variable selection methods in the literature can be applied to highly multivariate data sets, such as NIR spectroscopy, and to data sets consisting of process variables. Several studies can be found summarizing the best known and most widely used chemometric variable selection methods [14–17], highlighting Variable Importance in Projection (VIP) [18], Selectivity Ratio (SR) [19], interval PLS (iPLS) [20], and the Genetic Algorithms (GA) [21]. Some of the most recent developments in variable selection analysis on highly multivariate information are the works developed by Roger et al., [22] and Biaconlillo et al., [23]. They proposed two variable selection methods known as Covariance Selection (CovSel) and Sequential and Orthogonalized CovSel (SO-CovSel), respectively. The fundamental principle of both methods is the same. It consists in eliminating the collinear information linked to the independent variables (identified in each iteration of the method) that have a maximum covariance relationship with the response variable. The main difference between the proposed methods lies in the type of modelling applied, either single block (where CovSel is applied) or multi-block (where SO-CovSel is applied).

Variable selection on highly multivariate data has a wide range of applications and can be found in the fields of medicine [24], microbiology [25], food [26,27], and pharmaceuticals [28]. In the fuel domain, Villar et al. [29] evaluated the potential of the Martens uncertainty test, iPLS, and GA methods in selecting variables for the development of predictive models from Vis-NIR spectra acquired in marine diesel engine lubricating oil samples. They determined that when using the retained variables identified with the iPLS method (261/351), the online monitoring of the oil quality was improved.

Two more recent studies using variable selection to determine the quality of diesel fuel are worth mentioning. Nespeca et al. [30] evaluated the potential of three variable selection methods (Forward-iPLS, Backward iPLS, GA) applied on ATR-FTIR analysis performed on diesel samples to identify the variables that best described 8 properties of this fuel (density, flash point, sulfur content, cetane number, biodiesel content, and simulated distillation temperatures range to obtain 10%, 50% and 85% of sample distillate, T5, T50, and T85 respectively). With a similar performance of the three methods evaluated, the selection of variables

provided a more accurate description of the variables studied, except for T85. Shukla et al. [31] evaluated four variable selection methods (LASSO, correlation coefficient, Mallow's Cp criterion, Relative sensitivity ratio) applied on NIR spectra to improve the prediction of six diesel properties (cetane number, boiling point, freezing point, total aromatic content, viscosity and density), with the LASSO method showing the best performance.

Concerning the analysis of variable selection in a block of low univariate data, a study developed by Desboulets [32] describing the methods and its applications can be found. Among the methods that have shown better performance, the Least Absolute Shrinkage and Selection Operator (LASSO) [33], Recursive Feature Elimination (RFE)[34], and the sequential feature selection (SFS) [35], along with its variant the sequential floating feature selection method (SFFS) [36], can be highlighted.

The different studies for improving the estimation of final product properties and optimizing process monitoring reported in the literature and analyzed in this study can be classified into two approaches: fusion of analytical data and process variables, and variable selection. Taking into account this context and seeking to take advantage of the full potential of these two approaches, in this paper was analyzed a combined approach not reported in the literature related to fuel characterization. In summary, this study investigated the impact of different variable selection methods on the performance of data fusion models between the NIR spectra acquired on the total effluent obtained from the hydrocracking (HCK) process and the variables involved in this process for estimating the diesel cetane number. The work presented in this article was divided in three main steps, (i) the single modelling using each data block, (ii) the variable selection applied on the high and low multivariate blocks, and (iii) the data fusion. More information on the context of predicting the diesel cetane number from the analytical information of the total effluent can be found in [2].

# 2. Materials and methods

In this study, three different data sets were considered: HCK process information collected from laboratory analysis and process conditions, NIR spectra acquired the HCK total effluent samples, and cetane number measured on the diesel samples recovered from the distillation of the above-mentioned total effluent samples. This section details the source of these three data sets.

## 2.1 Hydrocracking process information

The first set of data consolidated was related to the HCK process operation. In this study, 53 variables, split in 3 groups, regarding the HCK process were collected from 64 experimental tests conducted at IFPEN HCK pilot plants in Solaize, France: (1) 23 are related to the quality of the feedstock, (2) 14 to the operational conditions, and (3) 16 to the total effluent characteristics. Table App 1 to Table App 3, provided as supplementary information in the appendix section, show the detail of each group and four statistical parameters calculated from the information gathered on these variables. It is important to highlight that in

Table App 2 the information regarding the catalytic system used was coded to respect the confidentiality agreements related to this type of information.

## 2.2 NIR spectra

From the experimental tests conducted in the HCK pilot plants at IFPEN under the different operating conditions summarized in Table App 2 and involving different catalytic systems, 64 total effluent samples were obtained by processing 13 different vacuum gasoils (VGO). Table App 3 shows the variability of the total effluent samples analyzed in this study. The NIR spectra employed in this study were acquired on the total effluent samples, which before acquisition were heated in a water bath at 60°C in a closed flask for one hour and then manually shaken to ensure their liquid state and homogeneity. The NIR spectra acquisition was conducted using a NIRS XDS Process Analyzer (Metrohm, Villebon - France) spectrometer, recording wavelengths from 800 - 2200 nm with a resolution of 0.5 nm. The immersion probe used was a reflectance Falcata Lab6 (Hellma GmbH & Co. KG, Müllheim – Germany) with an optical path fixed at 2 mm. Each final spectrum was the average of 32 scans performed on each sample. The software used with the spectrometer was VISION (Metrohm, Villebon - France).

Buendia Garcia et al. [2] validated and established that the region of the total effluent NIR spectrum providing the most descriptive information to predict the diesel cetane number is between 1110 and 2200 nm. Therefore, in this study, it was employed the same spectral region.

## 2.3 Cetane number

The diesel samples on which the cetane number was measured were recovered by atmospheric distillation of each of the 64 total effluents using the ASTM D2892-20 standard [37]. In other words, 64 diesel samples corresponding to the total effluents described previously were used in this study. The cetane number was measured on the diesel samples using an internal method developed in the IFPEN validated against the ASTM D613 standard method [38]. Table App 4 in the appendix section summarizes the general statistical information of this property, as well as of the density [39], and the simulated distillation temperatures range to obtain both 5% and 95% of sample distillate (Simulated Distillation T5 and T95) [40].

## 2.4 Modelling

For purposes of this study, two main data blocks were used; the NIR data block, identified in the study as Mx1 and having a size of [64 X 2180], and the process variables data block identified as Mx2 and having a size of [64 X 53]. In addition, the dependent variable matrix (My) was built with the respective measured diesel cetane number.

After consolidating each data block, they were split into the calibration and test sub-datasets using the Kennard and Stone (KS) algorithm [41] applied on the NIR block. The calibration sub-set (70% samples) was used both in model creation and cross-validation, and the independent test sub-set (30% samples) was used in the performance evaluation of the developed models. In this study, the calibration and test sub-datasets

used in developing all models were the same.

Before model calibration, the two data blocks were preprocessed either to correct and remove information not related to the chemical nature of the sample, as in the case of the NIR block, or to reduce the impact that the difference in scale of the variables may have on the data analysis, as in the case of the process data block. The NIR data block was preprocessed using the variable sorting for normalization (VSN) method [42], followed by the Savitzky-Golay first derivative using a 25-point window and a first-order polynomial (SavGol[25,1,1]) [43], and a column-wise data centering (mean center). Regarding the process variable data block, it should be highlighted that the source of information for each variable is diverse, resulting in different units of measurement and different scales. Therefore, to avoid that each variable impact on the model was governed by its scale rather than by its true contribution to the property description, this data block was scaled by subtracting the mean of each variable and then dividing each variable by its standard deviation.

The first step was to create single models for diesel cetane number estimation using each data block separately. The regression method used for the NIR data block was partial least squares (PLS), while the multiple linear regression (MLR) method was used for the process variable data block. The root mean squared errors of calibration (RMSEC), cross-validation (RMSECV), and prediction (RMSEP) were calculated for subsequent model performance evaluation. The RMSECV was used as the main criterion for defining the number of latent variables retained in the PLS model. These single models were defined as a reference for comparison with the models developed when applying the variable selection analysis.

The next step was implementing different variable selection methods on each data block. As summarized in Table App 5, seven different strategies were employed on the matrix Mx1 (NIR spectra). The RMSEC, RMSECV, and RMSEP, calculated on the different PLS models (MLR model for the CovSel method) developed from the datasets generated with each variable selection method implementation, were used to evaluate the performance of the different variable selection strategies. Moreover, the respective squared correlation coefficients between actual and predicted values ($r^2C$, $r^2CV$, $r^2P$) complemented the evaluation. Similarly, 11 different variable selection strategies, summarized in Table App 6, were applied on the matrix Mx2 (process variables). The best performing variable selection method was determined using the same evaluation criteria but calculated from a multiple linear regression (MLR) model.

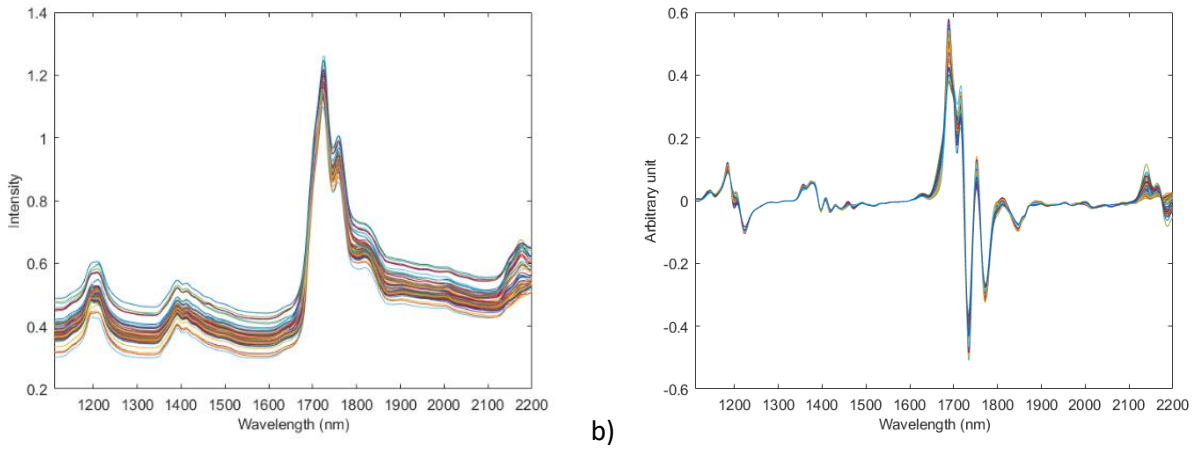
Once the most descriptive variables for each data block were identified using the adequate variable selection method, two data fusion models were developed using the high-level strategy [44] with the cetane number predicted from each block as the decision to fuse. The two models were developed using the data blocks with and without variable selection. Finally, the SO-CovSel method was evaluated. This method was proposed as a multi-block algorithm for variable selection and aimed to perform a more accurate selection of the descriptive variables of each data block by capturing their interaction. These three final models were

compared using the same six statistical criteria described before.

The models and variable selection analysis were performed using the PLS_Toolbox V.8.8 (Eigenvector Research Inc. Wenatchee, WA, USA), MATLAB V.2019b (The MathWorks, Inc., Natick, MA, USA), and Python V3.6.

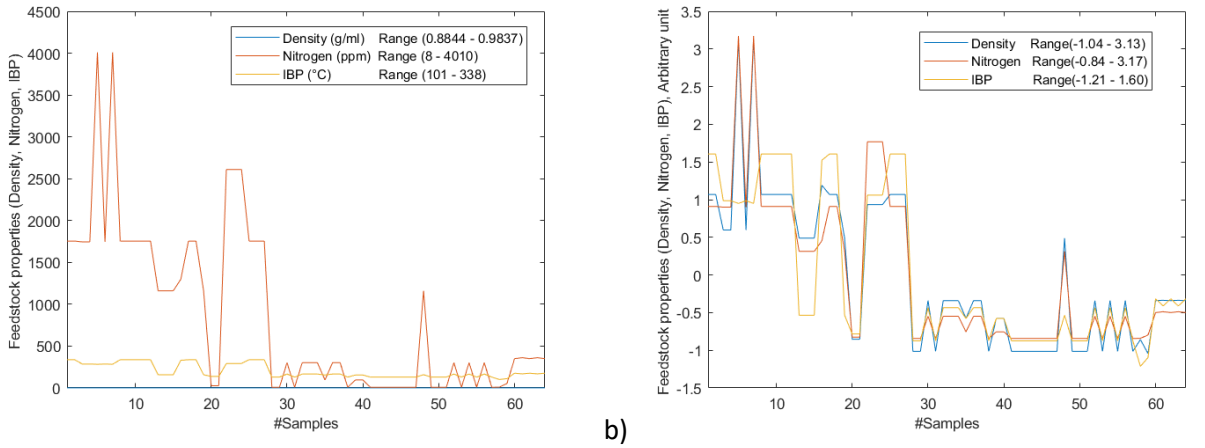
# 3. Results and discussion

Figure 1a shows the NIR raw spectra, while Figure 1b shows the preprocessed data. From Figure 1b, the four most influential zones containing the chemical information that best describes the behavior of the hydrocarbon molecules present in the samples can be observed [45,46]. Concerning the process information data set, an example of the difference in the scale of these variables that could impact the evaluation of the different variable methods is shown in Figure 2a, where three properties of the feedstock (density, initial boiling point (IBP), and nitrogen content) are compared. After preprocessing this data block (autoscale) it can be observed that the variables scale is comparable (Figure 2b).




a) b)

*Figure 1. NIR spectra comparison. a) Raw signal, b) preprocessed signal (VSN+SavGol[25,1,1])*




a) b)

*Figure 2. Process data comparison. a) Raw data, b) preprocessed data (auto scale)*

## 3.1 Single modelling

From the matrix Mx1, a PLS model of 5 latent variables (LVs) was developed, capturing 99% and 97% of the explained variance of the x and y matrices, respectively. In general, this model allows estimating the diesel cetane number reliably in a range of 35.1-67.9 from the NIR spectra of the total effluent, having errors close to and even lower than the reproducibility of the reference method (±3.6) (RMSEC = 1.6, RMSECV = 2.0, RMSEP = 1.9). Nevertheless, a detailed analysis of the model performance shows that four samples, three corresponding to the calibration set and 1 to the test set, are predicted outside the confidence limits presenting errors in the prediction up to 1.5 times the reproducibility of the reference method (see Figure 9a).

To confirm whether the four poorly predicted samples are outliers, the Q residual and Hotelling $T^2$ analysis applied to the calibration and test data sets was conducted. Figure 3 shows that all four samples (red color) are within the thresholds of the two analyses confirming their non-outlier condition. After discarding the outlier character of these samples in the model, a more detailed analysis was performed on the operating conditions for obtaining the total effluent and the acquisition of the NIR spectra, resulting in not finding any particular explanation for the anomalous behavior of these samples. This suggests that the model may not fully capture the descriptive information of the samples NIR spectra to predict them correctly.

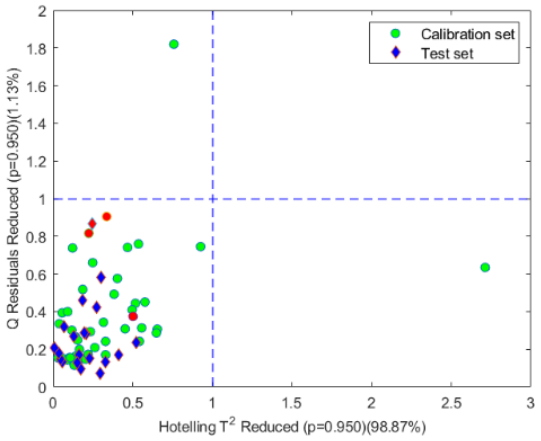



*Figure 3. Reduced Q residual and Hotelling $T^2$ analysis using a 5 LVs PLS model*

The second single model generated was from the Mx2 matrix. The MLR model obtained presents an RMSEP (1.8) close to the one obtained with the PLS model using the Mx1. However, the RMSEP/RMSEC ratio (4.5) is quite high, indicating the possible overfitting of the model. This can be seen in Figure 9b, where an almost perfect prediction is observed for the calibration dataset ($r^2C$ = 0.998), while the prediction of the test set samples presents a higher scatter ($r^2P$ = 0.955). The notion of a possible model overtraining is strengthened when the RMSECV (3.8) is included in the analysis. For this reason, the use of this model should be done cautiously.

An interesting finding in the analysis of this MLR model is that one of the samples of the test dataset that the single NIR model predicted outside the reproducibility limits is now predicted within these limits. However, the MLR model predicts a sample outside the same limits correctly estimated in the individual NIR model. This leads to assume that the information contained in the two data sets could complement each other for better accuracy in the diesel cetane number estimation.

The final comparison table reports detailed information of the two single models developed (see Table 4).

## 3.2 Variable selection analysis on NIR data block

The different variable selection methods, along with their application parameters summarized in Table App 5, were applied to the calibration data set of the Mx1. Table 1 shows the statistical parameters calculated for evaluating each model performance, while the Figure 4 summarizes the number of variables selected by each method. In addition, as described in the materials and methods section, PLS models were developed from the different groups of selected variables (MLR for the CovSel method) and then applied on the test data set to evaluate the effectiveness of each selection method in defining the most descriptive variables. The evaluation parameters of these models (errors and squared correlation coefficients) are also summarized in Table 1.

*Table 1. Results summary of variable selection methods applied on NIR data block*

| | Models description | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Model** | **LVs** | **RMSEC** | **$r^2C$** | **RMSECV** | **$r^2CV$** | **RMSEP** | **$r^2P$** |
| VIP | PLS | 5 | 1.7 | 0.986 | 2.0 | 0.955 | 1.9 | 0.951 |
| SR | | 5 | 1.7 | 0.968 | 2.0 | 0.955 | 1.9 | 0.951 |
| GA | | 5 | 1.4 | 0.976 | 1.8 | 0.962 | 1.9 | 0.950 |
| F-iPLS | | 8 | 1.3 | 0.981 | 1.8 | 0.962 | 2.0 | 0.951 |
| B-iPLS | | 5 | 1.6 | 0.970 | 1.9 | 0.960 | 1.9 | 0.954 |
| rPLS | | 4 | 1.6 | 0.971 | 1.8 | 0.964 | 1.9 | 0.956 |
| CovSel | MLR | - | 0.8 | 0.993 | 1.5 | 0.975 | 1.4 | 0.976 |

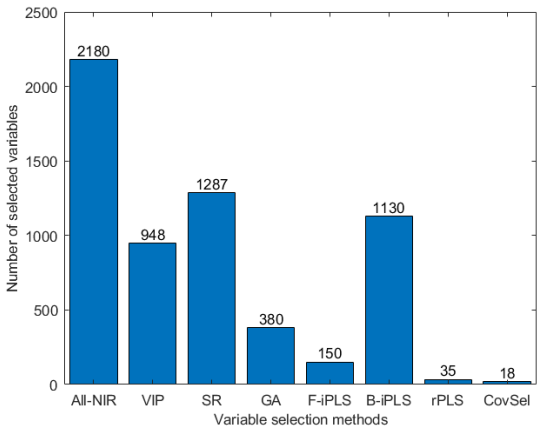



*Figure 4. Comparison of the number of selected variables in the NIR data block by each method*

Figure 5 graphically shows the performance comparison of the different models developed. Compared to the single model using all NIR variables, this figure indicates that the performance of the different PLS models obtained with the selected variables is quite close regarding the RMSE's and their respective $r^2$'s. Nevertheless, the reduction in the number of variables to be used is evident, especially with the GA, F-iPLS, and rPLS methods (See Figure 4). Although a significant improvement in the estimation of the studied property was not achieved, the selection of variables helps to identify the wavelengths of the NIR spectra acquired on the total effluent that best describe the behavior of the diesel cetane number, which could lead to cost optimization in acquiring the data.

Compared to the analyzed methods for selecting variables on the present dataset, CovSel outperforms the others, since, with an optimal selection of the number of variables to be used (18), it is the one that exhibits the lowest RMSECV and RMSEP, reflected in the improvement of the $r^2$CV and $r^2$P. Figure 9c shows the improvement in the estimation of the diesel cetane number by correctly predicting the three samples of the calibration dataset that the NIR model developed using all variables estimated outside the reproducibility limits of the reference method. However, this improvement is not reflected in the test set sample initially estimated outside these limits, corroborating the possible lack of descriptive information that the PLS model cannot capture from the NIR spectra.

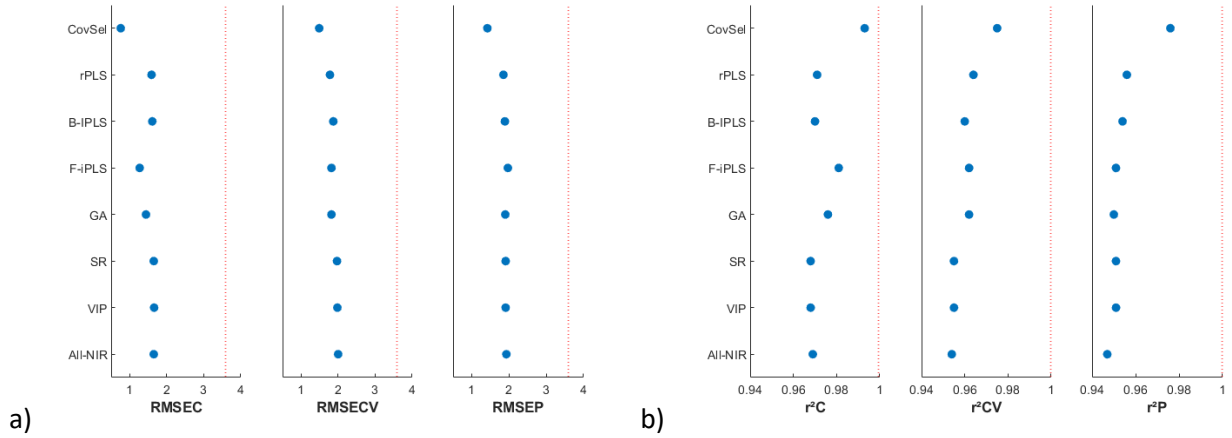



*Figure 5. Performance comparison of variable selection methods applied on NIR data block. a) model errors, red dotted line ⍰ reproducibility of the reference method (3.6). b) model squared correlation coefficients*

In order to evaluate the coherence of the variables selected by the CovSel method, they were identified along the NIR spectrum acquired on one of the total effluent samples used in this study. From Figure 6, it can be observed that 12 of the selected variables are located in the spectral region corresponding to the first overtone bands of -CH stretch in -CH2 and -CH3 (1600-1900 nm), while the other six variables are located in the region corresponding to the combination absorption bands of -CH stretching bonds and C=C stretching bonds in the aromatic ring (2100-2200 nm). Analyzing the MLR model in detail, it was found that the group of the 12 variables identified in the region between 1600-1900 nm contributes 80% of the weight in the estimation of the diesel cetane number. Therefore, if it is considered that the amount and interaction of

linear hydrocarbons compounds ([CH3-(CH2)n-CH3]) directly impacts the behavior of the cetane number 221, the variables selected by the CovSel method are meaningful and validate their consistency.

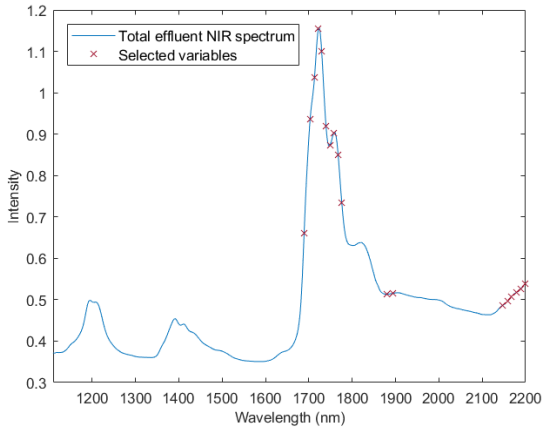



*Figure 6. CovSel selected variables identification over an average NIR spectrum acquired on a total effluent sample*

## 3.3 Variable selection analysis on process variables data block

For the variable selection analysis applied in Mx2, the 11 methods summarized in Table App 6 were used. The performance of the different methods evaluated was assessed based on the RMSE's and $r^2$'s calculated for the different MLR models developed from the groups of the selected variables. Table 2 shows the results obtained in this evaluation, while Figure 7 shows the number of variables selected by each method.

*Table 2 Results summary of variable selection methods applied on process variables data block*

| | Models description | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **RMSEC** | **$r^2$C** | **RMSECV** | **$r^2$CV** | **RMSEP** | **$r^2$P** |
| VIP | 0.5 | 0.997 | 3.7 | 0.670 | 2.0 | 0.946 |
| SR | 0.4 | 0.998 | 3.8 | 0.862 | 1.8 | 0.955 |
| LASSO | 1.8 | 0.964 | 2.5 | 0.927 | 2.0 | 0.944 |
| GA | 1.2 | 0.983 | 1.6 | 0.971 | 2.2 | 0.929 |
| RFE | 1.3 | 0.981 | 1.7 | 0.910 | 1.9 | 0.960 |
| XGBoost_FS | 1.6 | 0.971 | 2.4 | 0.933 | 1.7 | 0.960 |
| SFS | 2.5 | 0.931 | 3.4 | 0.868 | 2.5 | 0.935 |
| SBS | 1.5 | 0.974 | 3.2 | 0.890 | 1.8 | 0.963 |
| SFFFS | 1.1 | 0.987 | 1.7 | 0.965 | 1.5 | 0.970 |
| SBFFS | 1.4 | 0.978 | 1.7 | 0.969 | 1.1 | 0.984 |
| CovSel | 0.5 | 0.997 | 2.9 | 0.914 | 1.6 | 0.968 |

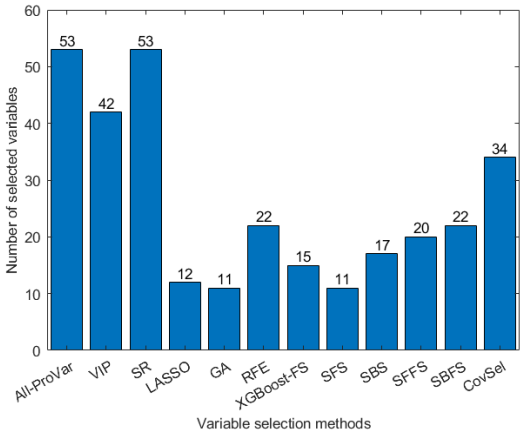



*Figure 7. Comparison of the number of selected variables in the process information data block by each method*

Analogously to the analysis done on the NIR block, Figure 8 shows the performance of the different MLR models developed according to the RMSEs and $r^2$s. The first issue that stands out is the ineffectiveness of the VIP and SR methods in selecting variables when analyzing this type of data, as they provide a small reduction of variables employed, or none at all as in the case of SR. The other methods show greater efficiency in the variable selection procedure, reflected in the improved stability of the errors, particularly the RMSECV. Out of these methods, GA, RFE, SFFFS, and SBFFS stand out. With 22 selected variables, the SBFFS method has the best performance. Besides presenting the highest accuracy in estimating the diesel cetane number, the variables identified with this method allow having the lowest RMSEP and the highest $R^2P$.

Interestingly, the high efficiency in variable selection shown by the CovSel method on the NIR data block is not observed when applied to the process variable data block. This could be attributed to the premise that the principle of this method was based on highly multivariate data, limiting its performance on less multivariate data.

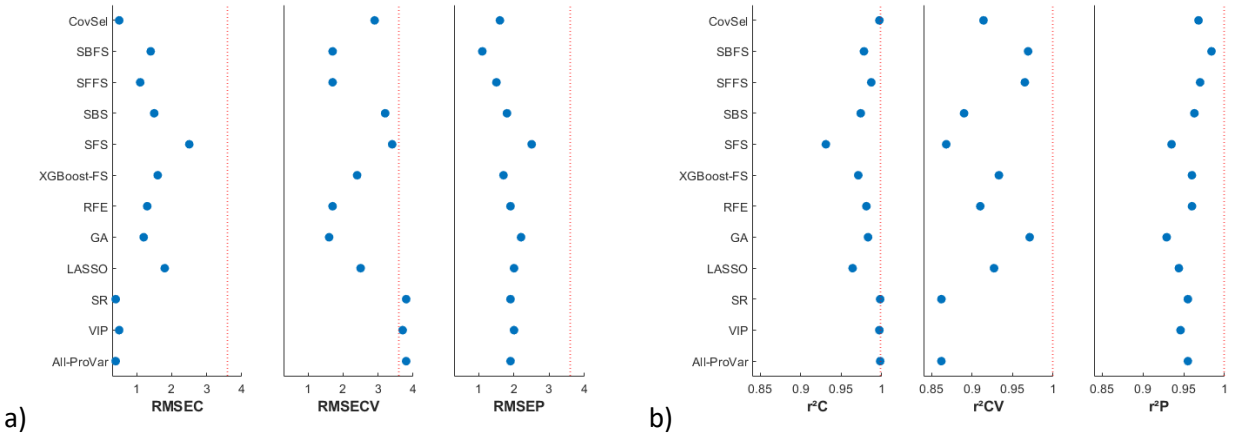



*Figure 8. Performance comparison of variable selection methods applied on process variables data block. a) model errors, red dotted line ⍰ reproducibility of the reference method (3.6). b) model squared correlation coefficients*

When thoroughly analyzing the performance of the MLR model developed from the 22 variables selected by the SBFFS method, it was possible to verify the improvement in the estimation of the cetane number in new samples by predicting within the reproducibility limits all the samples corresponding to the test data set. However, one sample from the calibration data set is predicted outside these limits (See Figure 9d).

The variables selected by the SBFFS method, summarized in Table 3, were analyzed to determine their relationship with the diesel cetane number. It should be emphasized that 50% of the selected variables correspond to the total effluent. These variables can be summarized as density, refractive index, and simulated distillation. An analysis of the MLR model determined that these 11 variables represent 67% of the weight in the model, being the density and the refractive index the most significant contributors. A general analysis can illustrate their consistent relationship with the cetane number. Firstly, the simulated distillation provides information about the diesel volume to be recovered (also known as diesel conversion). Secondly, the density and refractive index supply information about the interaction of the linear chains and aromatic compounds contained in the sample [47,48].

Regarding the variables selected concerning the operating conditions, they contribute to 18% of the weight in the model, being the reaction temperature in the HDT reactor and the catalyst of the HCK stage the most important contributors to the explanation of the diesel cetane number behavior. Once again, the coherence between the selected variables and the studied property can be verified since is in the HCK stage where the reactions for cracking the long-chain molecules occur leading to the generation of hydrogenated linear chains of smaller size. These short-chain molecules generated impact the cetane number directly. Finally, the remaining 15% of the model weight is given by the quality of the feedstock, where the paraffinic carbon content and the simulated distillation are the most informative variables.

## 3.4 Data fusion modelling

Aiming to evaluate the contribution of the variable selection in the synergic use of the two data blocks for estimating the cetane number, two data fusion models were generated using the high-level strategy with the cetane number predicted from each block as the decision to fuse. Table 4 presents the description of the models developed, while Figure 9 shows the comparative analysis of these models.

The first model was calibrated using all the variables of the two data blocks to determine the data fusion decision. In addition, the possible combinations of latent variables in the PLS model from the Mx1 matrix with the MLR model from the Mx2 matrix were tested in this first model. Compared to the individual model of the NIR block using all variables (information summarized in Table 4), this first data fusion model reduces the RMSECV and RMSEP by about 50% and 20%, respectively. Concerning the individual model of the process variable block, the errors are improved by about 74% and 15%. The resulting performance improvement of the model due to data fusion is also reflected in the $r^2C$ and $r^2P$. However, although a significant reduction in

the prediction error of the samples corresponding to the test data set is achieved, a sample is still not estimated within the reproducibility limits (see Figure 9e). This can be attributed to the use of redundant or uninformative information in the development of the model, a limitation that can be addressed with the proper selection of each data block descriptors.

*Table 3. Results summary of variable selection methods applied on process variables data block*

| Process group source | Variable |
|---|---|
| Feedstock | Nitrogen (ppm) |
| | Paraffinic carbon (%wt) |
| | Initial Boiling Point (°C) |
| | SimDis T40(°C) |
| | SimDis T70(°C) |
| | SimDis T95(°C) |
| | Final Boiling Point (°C) |
| Operating conditions | Pressure (bar) |
| | Temperature R1 (°C) |
| | HDT Catalyst . Parameter 1 |
| | HCK Catalyst. Parameter 2 |
| Total effluent | Density (g/ml) |
| | Refractive Index |
| | SimDis T5(°C) |
| | SimDis T10(°C) |
| | SimDis T30(°C) |
| | SimDis T50(°C) |
| | SimDis T60(°C) |
| | SimDis T80(°C) |
| | SimDis T90(°C) |
| | SimDis T95(°C) |
| | Final Boiling Point (°C) |

*Table 4. Results summary of single and data fusion models*

| Model type | Data block | Variable selection method | Variables used | Model | RMSEC | $r^2C$ | RMSECV | $r^2CV$ | RMSEP | $r^2P$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Single | NIR | - | 2180 | PLS (5 LV's) | 1.6 | 0.969 | 2.0 | 0.954 | 1.9 | 0.947 |
| | | CovSel | 18 | MLR | 0.8 | 0.993 | 1.5 | 0.975 | 1.4 | 0.976 |
| | Process variables (ProVar) | - | 53 | | 0.5 | 0.998 | 3.8 | 0.862 | 1.8 | 0.955 |
| | | SBFS | 22 | | 1.4 | 0.978 | 1.7 | 0.969 | 1.1 | 0.984 |
| Data fusion | NIR + Provar | - | 2* [2180+53] | MLR** [PLS+MLR] | 1.0 | 0.989 | 1.0 | 0.988 | 1.5 | 0.972 |
| | | CovSel-SBFS | 2* [18+22] | MLR** [MLR+MLR] | 0.6 | 0.981 | 0.7 | 0.995 | 0.9 | 0.988 |

* The two variables are the cetane number predicted from each data block as the decision to fuse

** The MLR model was developed from the cetane number predicted from each data block

The second data fusion model was developed using only the variables selected in each data block, i.e., 18 variables from the NIR block identified with the CovSel method and 22 descriptors from the process information block defined with the SBFS method. This data fusion model presents an outstanding performance by reducing around 67% and 47% the RMSECV and RMSEP compared to the individual NIR model, estimating all the samples of both the calibration and test datasets within the reproducibility limits (see Figure 9f). In addition, the model presents higher stability regarding the RMSE's and the $r^2$'s. Analyzing the performance of the two data fusion models, the importance of the identification and selection of variables that best describe the studied property is evident.

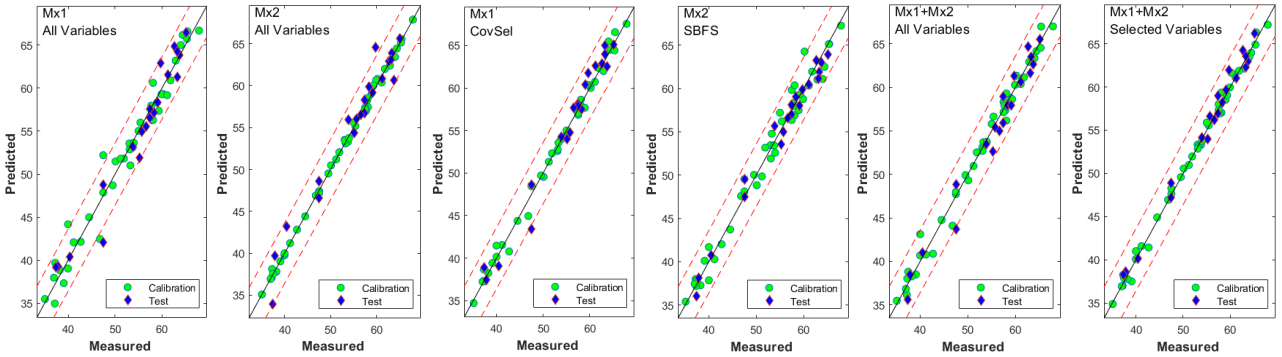



*Figure 9. Parity plot for model performance comparison. a) PLS model using all variables from Mx1, b) MLR model using all variables from Mx2, c) MLR model using selected variables from Mx1, d) MLR model using selected variables from Mx2, e) Data fusion model using all variables from Mx1 & Mx2, f) Data fusion model using selected variables from Mx1 & Mx2. Circle shape □ calibration samples, diamond shape □ test samples*

The results shown in this study suggest that the selection of most descriptive variables and the method used to do so have a significant impact in optimizing the performance of the data fusion model. Therefore, it was desired to evaluate the SO-CovSel method, whose objective is focused on the variable selection in multi-block modelling. Using a MATLAB script developed by Federico Marini [23] which applying the SO-CovSel concept evaluates different combinations of variables between the two data blocks selecting the best combination based on the RMSECV of the final orthogonalized model, it was possible to construct the error plot shown in Figure 10. In this figure, it can be observed that the best-performing model is obtained using seven variables in total (5 from the Mx1 and 2 from the Mx2) having RMSEC, RMSECV, and RMSEP of 1.4, 2.3 and 3.1, respectively. Compared to the individual and data fusion models with variable selection, the model developed by the SO-CovSel method shows no improvement in estimating the diesel cetane number, having higher RMSEP. However, it should be noted that the RMSE's are still below the reproducibility limit of the reference method using only seven variables. In the scenario where the main objective of the variable selection was the costs optimization in the information acquisition process rather than the improvement of the estimation of the studied property, the SO-CovSel presents great effectiveness in identifying the descriptors allowing an estimation with errors close to the reference method. Furthermore, the possible limited performance of the SO-CovSel method shown in this study could be attributed to the low multivariate nature of the Mx2, as previously evidenced by the application of the CovSel method on this data block.

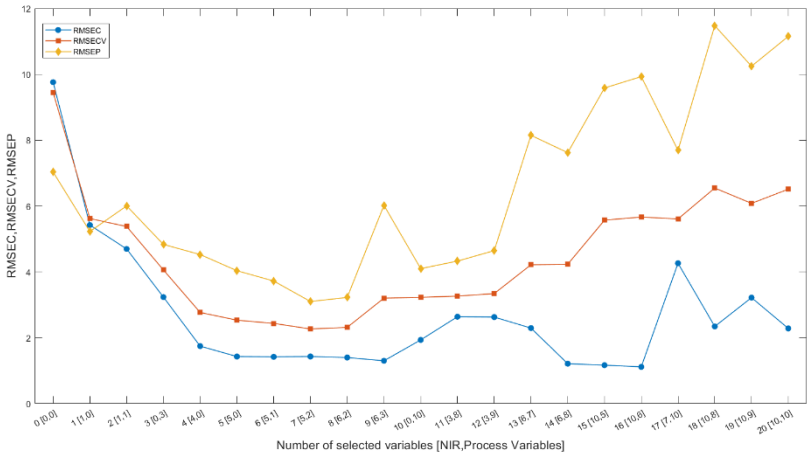



*Figure 10. SO-CovSel model RMSE's trend*

# Conclusions

This study evaluated the contribution of variable selection analysis in improving NIR and process information data fusion model performance in predicting diesel cetane number. The evaluation was conducted applying seven variable selection methods on the NIR block and eleven methods on the process variable block before data fusion modelling. In addition, the potential of the SO-CovSel multi-block variable selection method was also analyzed.

The data fusion modelling using all variables from each data block improves the estimation of the diesel cetane number compared to single models. However, using variable selection analysis before data fusion significantly improves the estimation of this property and leads to greater model stability regarding the RMSE's and $r^2$'s.

The CovSel method for variable selection exhibited a remarkable performance when applied to the NIR data block. Compared to the other methods evaluated on this block, CovSel achieved the highest optimization in the number of identified variables that best explained the property behavior, resulting in a better estimation. However, the performance of this method was not as good when applied to the data block of process variables. Considering that the CovSel method was developed primarily for highly multivariate data, its effectiveness could be affected by a less multivariate data set, such as the process variables information. For this dataset, the method that presented the best performance was the SBFS.

The multi-block variable selection method SO-CovSel did not improve the estimation of the studied property. Adopting the CovSel method as the underlying multi-block modelling algorithm, the limited performance of SO-CovSel could be attributed to the low multivariate nature of the process variable data block. However, when compared to the reference method used for cetane number measurement, the SO-CovSel method provides estimates with errors close to its reproducibility using a minimum number of variables. This could potentially lead to an optimization of data acquisition costs.

The results from this study highlight the great impact that variable selection can have in improving the estimation of diesel cetane number and understanding the parameters affecting the behavior of this property. Furthermore, the advantage offered by this analysis could be resulted not only in having a more accurate prediction but also in optimizing the process analysis and the resources required for data acquisition.

# Acknowledgments

The authors would like to thank IFP Energies Nouvelles for providing the total effluent samples from the HCK process reactors, the facilities for the distillation to obtain the diesel samples, and the facilities for spectra acquisition and data analysis. Thanks also to Axel One Analysis for providing the probe for the NIR spectra acquisition. Special thanks to Federico Marini for the MATLAB script used in the analysis of the SO-CovSel method.

# CRediT authorship contribution statement

**J. Buendia-Garcia**: Conceptualization, Data curation, Writing - original draft. **J. Gornay:** Conceptualization, Writing - original draft. **M. Lacoue-Negre:** Conceptualization, Writing - original draft. **S. Mas-Garcia:** Writing - original draft. **R. Bendoula:** Writing - original draft, **J.M Roger:** Conceptualization, Writing - original draft

# Declaration of conflicting interests

The Author(s) declare(s) that there is no conflict of interest

# REFERENCES

[1] Moro MK, dos Santos FD, Folli GS, Romão W, Filgueiras PR. A review of chemometrics models to predict crude oil properties from nuclear magnetic resonance and infrared spectroscopy. Fuel 2021;303. https://doi.org/10.1016/j.fuel.2021.121283.

[2] Buendia Garcia J, Lacoue-Negre M, Gornay J, Mas Garcia S, Bendoula R, Roger JM. Diesel cetane number estimation from NIR spectra of hydrocracking total effluent. Submitted to Fuel 2022, 2022.

[3] Killner MH, Rohwedder JJ, Pasquini C. A PLS regression model using NIR spectroscopy for on-line monitoring of the biodiesel production reaction. Fuel 2011;90(11):3268–73. https://doi.org/10.1016/j.fuel.2011.06.025.

[4] He K, Qian F, Cheng H, Du W. A novel adaptive algorithm with near-infrared spectroscopy and its application in online gasoline blending processes. Chemometrics and Intelligent Laboratory Systems 2015;140:117–25. https://doi.org/10.1016/j.chemolab.2014.11.006.

[5] de Oliveira RR, Pedroza RH, Sousa AO, Lima KM, de. Juan A. Process modelling and control applied to real-time monitoring of. Analytica Chimica Acta 2017, 2017:41–53.

[6] Lambert D, Saint-Martin C, Benkhelil K, Ribero B, Valleur M. Advanced crude management by NIR spectroscopy combined with topology modelling. Hydrocarbon Processing 2019.

[7] AlGhazzawi A, Lennox B. Monitoring a complex refining process using multivariate statistics. Control Engineering Practice 2008;16(3):294–307. https://doi.org/10.1016/j.conengprac.2007.04.014.

[8] Smolinska A, Engel J, Szymanska E, Buydens L, Blanchet L. General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences. Data Fusion Methodology and Applications 2019:51–79. https://doi.org/10.1016/B978-0-444-63984-4.00003-X.

[9] Moro MK, Neto ÁC, Lacerda V, Romão W, Chinelatto LS, Castro EV et al. FTIR, 1H and 13C NMR data fusion to predict crude oils properties. Fuel 2020;263:116721. https://doi.org/10.1016/j.fuel.2019.116721.

[10] de Aguiar LM, Galvan D, Bona E, Colnago LA, Killner MHM. Data fusion of middle-resolution NMR spectroscopy and low-field relaxometry using the Common Dimensions Analysis (ComDim) to monitor diesel

fuel adulteration. Talanta 2022;236:122838. https://doi.org/10.1016/j.talanta.2021.122838.
[11] Mishra P, Marini F, Biancolillo A, Roger J-M. Improved prediction of fuel properties with near-infrared spectroscopy using a complementary sequential fusion of scatter correction techniques. Talanta 2021;223(Pt 1):121693. https://doi.org/10.1016/j.talanta.2020.121693.
[12] de Oliveira RR, Avila C, Bourne R, Muller F, de. Juan A. Data fusion strategies to combine sensor and multivariate model outputs for multivariate statistical process control. Anal Bioanal Chem 2020;412(9):2151–63. https://doi.org/10.1007/s00216-020-02404-2.
[13] Strani L, Mantovani E, Bonacini F, Marini F, Cocchi M. Fusing NIR and Process Sensors Data for Polymer Production Monitoring. Front Chem 2021;9:748723. https://doi.org/10.3389/fchem.2021.748723.
[14] Andersen CM, Bro R. Variable selection in regression-a tutorial. Journal of Chemometrics 2010;24(11-12):728–37. https://doi.org/10.1002/cem.1360.
[15] Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M. Variables selection methods in near-infrared spectroscopy. Analytica Chimica Acta 2010;667(1-2):14–32. https://doi.org/10.1016/j.aca.2010.03.048.
[16] Anzanello MJ, Fogliatto FS. A review of recent variable selection methods in industrial and chemometrics applications. EJIE 2014;8(5):619. https://doi.org/10.1504/EJIE.2014.065731.
[17] de Araújo Gomes A, Azcarate SM, Diniz PHGD, de-Sousa Fernandes DD, Veras G. Variable selection in the chemometric treatment of food data: A tutorial review. Food Chem 2022;370:131072. https://doi.org/10.1016/j.foodchem.2021.131072.
[18] Wold S, Johansson A, Cochi M. PLS-partial least squares projections to latent structures; 1993.
[19] Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometrics and Intelligent Laboratory Systems 2009;95(1):35–48. https://doi.org/10.1016/j.chemolab.2008.08.004.
[20] Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval Partial Least-Squares Regression (i PLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Appl Spectrosc 2000;54(3):413–9. https://doi.org/10.1366/0003702001949500.
[21] Goldberg DE. Genetic algorithms in search, optimization, and machine learning. Reading (Mass.) etc.: Addison-Wesley publishing company; 1989.
[22] Roger JM, Palagos B, Bertrand D, Fernandez-Ahumada E. CovSel: Variable selection for highly multivariate and multi-response calibration. Chemometrics and Intelligent Laboratory Systems 2011;106(2):216–23. https://doi.org/10.1016/j.chemolab.2010.10.003.
[23] Biancolillo A, Marini F, Roger J-M. SO-CovSel: A novel method for variable selection in a multiblock framework. J. Chemometrics 2020;34(2). https://doi.org/10.1002/cem.3120.
[24] Jesus J, Araujo D, Canuto A. Fusion Approaches of Feature Selection Algorithms for Classification Problems 2016:379–84. https://doi.org/10.1109/BRACIS.2016.075.
[25] de Sousa Marques A, de Melo MCN, Cidral TA, Gomes de Lima, Kássio Michell. Feature selection strategies for identification of Staphylococcus aureus recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: a case study. J Microbiol Methods 2014;98:26–30. https://doi.org/10.1016/j.mimet.2013.12.015.
[26] Valderrama P, Braga JWB, Poppi RJ. Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. J Agric Food Chem 2007;55(21):8331–8. https://doi.org/10.1021/jf071538s.
[27] Murphy TB, Dean N, Raftery AE. Variable Selection and Updating In Model-Based Discriminant Analysis for High Dimensional Data with Food Authenticity Applications. Ann Appl Stat 2010;4(1):396–421. https://doi.org/10.1214/09-AOAS279.
[28] Cui Y, Song X, Chuang K, Venkatramani C, Lee S, Gallegos G et al. Variable selection in multivariate modelling of drug product formula and manufacturing process. J Pharm Sci 2012;101(12):4597–607. https://doi.org/10.1002/jps.23322.
[29] Villar A, Fernández S, Gorritxategi E, Ciria JI, Fernández LA. Optimization of the multivariate calibration of a Vis–NIR sensor for the on-line monitoring of marine diesel engine lubricating oil by variable selection methods. Chemometrics and Intelligent Laboratory Systems 2013, 2013:68–75.
[30] Nespeca MG, Hatanaka RR, Flumignan DL, de. Oliveira JE. Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis. Journal of Analytical Methods in Chemistry 2018:1–10.

https://doi.org/10.1155/2018/1795624.
[31] Shukla A, Bhatt H, Pani AK (eds.). Variable selection and modelling from NIR spectra data: A case study of diesel quality prediction using LASSO and Regression Tree; 2020.
[32] Desboulets L. A Review on Variable Selection in Regression Analysis. Econometrics 2018;6(4):45. https://doi.org/10.3390/econometrics6040045.
[33] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996, 1996:267–88; Available from: http://www.jstor.org/stable/2346178.
[34] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines 2002;46:389–422. https://doi.org/10.1023/A:1012487302797.
[35] Last M, Kandel A, Maimon O. Information-theoretic algorithm for feature selection. Pattern Recognition Letters 2001;22(6-7):799–811. https://doi.org/10.1016/S0167-8655(01)00019-8.
[36] Nakariyakul S, Casasent DP. An improvement on floating search algorithms for feature subset selection. Pattern Recognition 2009;42(9):1932–40. https://doi.org/10.1016/j.patcog.2008.11.018.
[37] ASTM D 2892-20. Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column). West Conshohocken, PA: ASTM International; 2020. https://doi.org/10.1520/D2892-20.
[38] ASTM D613-01. Test Method for Cetane Number of Diesel Fuel Oil. West Conshohocken, PA: ASTM International; 2001. https://doi.org/10.1520/D0613-01.
[39] ASTM D1218 - 12. Standard Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids; Available from: https://www.astm.org/Standards/D1218.htm.
[40] ASTM D2887 - 19ae1. Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography; Available from: https://www.astm.org/Standards/D2887.htm.
[41] Kennard RW, Stone LA. Computer Aided Design of Experiments. Technometrics 1969;11(1):137–48. https://doi.org/10.1080/00401706.1969.10490666.
[42] Rabatel G, Marini F, Walczak B, Roger J-M. VSN: Variable sorting for normalization. Journal of Chemometrics 2020;34(2):205. https://doi.org/10.1002/cem.3164.
[43] Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 1964;36. https://doi.org/10.1021/ac60214a047.
[44] Ballabio D, Todeschini R, Consonni V. Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data. In: Cocchi M, editor. Data fusion methodology and applications. Amsterdam: Elsevier; 2019, p. 129–155.
[45] Yalvac ED, Seasholtz MB, Crouch SR. Evaluation of Fourier Transform Near-Infrared for the Simultaneous Analysis of Light Alkene Mixtures. Appl. Spectrosc., AS 1997;51(9):1303–10. https://doi.org/10.1366/0003702971942303.
[46] Kelly JJ, Callis JB. Nondestructive analytical procedure for simultaneous estimation of the major classes of hydrocarbon constituents of finished gasolines. Anal. Chem. 1990;62(14):1444–51. https://doi.org/10.1021/ac00213a019.
[47] Creton B, Dartiguelongue C, de-Bruin T, Toulhoat H. Prediction of the Cetane Number of Diesel Compounds Using the Quantitative Structure Property Relationship. Energy Fuels 2010;24(10):5396–403. https://doi.org/10.1021/ef1008456.
[48] Butnaru I, Bruma M, Kopnick T, Stumpe J. Influence of Chemical Structure on the Refractive Index of Imide-Type Polymers. Macromol. Chem. Phys. 2013;214(21):2454–64. https://doi.org/10.1002/macp.201300309.
[49] ASTM D445-97. Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity). West Conshohocken, PA: ASTM International.
[50] ISO 20846. Petroleum products — Determination of sulfur content of automotive fuels — Ultraviolet fluorescence method: ISO International; 2011.
[51] ASTM D5291. Standard Test Methods for Instrumental Determination of Carbon, Hydrogen, and Nitrogen in Petroleum Products and Lubricants. West Conshohocken, PA: ASTM International; 2007.
[52] ASTM D 3238 - 95. Standard Test Method for Calculation of Carbon Distribution and Structural Group Analysis of Petroleum Oils by the n-d-M Method. West Conshohocken, PA: ASTM International; 2000.
[53] ASTM D 7213-15. Standard Test Method for Boiling Range Distribution of Petroleum Distillates in the Boiling Range from 100 °C to 615 °C by Gas Chromatography. West Conshohocken, PA: ASTM International; 2015.

[54] Rinnan Å, Andersson M, Ridder C, Engelsen SB. Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS. J. Chemometrics 2014;28(5):439–47. https://doi.org/10.1002/cem.2582.
[55] Chen T, Guestrin C. XGBoost 2016:785–94. https://doi.org/10.1145/2939672.2939785.

*Table App 4.1. Summary of the variability of the properties measured on the HCK process feedstocks*

| Property | Standard Method | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|---|
| Density (g/mL) | ASTM S1218-12 [39] | 0.85 | 0.98 | 0.91 | 0.027 |
| Refractive index | | 1.44 | 1.53 | 1.48 | 0.020 |
| Viscosity @ 70°C (cSt) | ASTM D445-97 [49] | 3.0 | 20.6 | 8.2 | 2.96 |
| Viscosity @ 100°C (cSt) | | 6.1 | 75.9 | 20.8 | 11.22 |
| Sulphur (wt%) | ISO 20846, 2011 [50] | 7E-04 | 3.5 | 1.1 | 1.08 |
| Nitrogen (ppm) | ASTM D5291, 2007 [51] | 2.2 | 4825.0 | 1160.9 | 1143.79 |
| Hydrogen (wt%) | | 10.6 | 13.8 | 12.5 | 0.74 |
| Aromatic Carbon (wt%) | ASTM D3238-95, 2000 [52] | 4.4 | 36.0 | 15.5 | 7.45 |
| Paraffinic Carbon (wt%) | | 43.1 | 71.6 | 57.1 | 5.33 |
| Naphtenic Carbon (wt%) | | 8.3 | 56.9 | 28.1 | 9.69 |
| SimDis IBP(°C) | ASTM D7213-15,2015 [53] | 68 | 358 | 228 | 91.1 |
| SimDis T5(°C) | | 122 | 404 | 324 | 56.5 |
| SimDis T10(°C) | | 160 | 415 | 357 | 43.5 |
| SimDis T20(°C) | | 216 | 436 | 391 | 33.6 |
| SimDis T30(°C) | | 269 | 450 | 412 | 28.4 |
| SimDis T40(°C) | | 323 | 466 | 429 | 25.0 |
| SimDis T50(°C) | | 369 | 480 | 445 | 21.8 |
| SimDis T60(°C) | | 390 | 498 | 462 | 20.4 |
| SimDis T70(°C) | | 409 | 515 | 480 | 19.2 |
| SimDis T80(°C) | | 433 | 537 | 501 | 19.2 |
| SimDis T90(°C) | | 466 | 563 | 529 | 18.7 |
| SimDis T95(°C) | | 493 | 606 | 552 | 17.0 |
| SimDis FBP(°C) | | 558 | 686 | 611 | 17.9 |

*SimDis = Simulated Distillation*
*IBP = Initial Boiling Point*
*T5-T95 = Temperature to recover 5%-95% of distilled sample*
*FBP = Final Boiling Point*

*Table App 4.2. Summary of the variability of the HCK process operating conditions*

| Parameter | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|
| Pressure (bar) | 30 | 160 | 121 | 29.67 |
| Temperature R1 (°C) | 350 | 415 | 385 | 14.6 |
| Temperature R2 (°C) | 370 | 420 | 392 | 13.1 |
| LHSV ($s^{-1}$) | 0.4 | 4.0 | 1.6 | 0.97 |
| HDT catalyst | Parameters CHDT1, CHDT2, CHDT3, CHDT4, CHDT5 | | | |
| HCK catalyst | Parameters CHCK1, CHCK2, CHCK3, CHCK4, CHCK5 | | | |

*LHSV = Liquid Hourly Space Velocity*
*HDT = Hydrotreating*
*HDT = Hydrocracking*

*Table App 4.3. Summary of the variability of the HCK process total effluents*

| Property | Standard method | Minimum | Maximum | Mean | Standard deviation |
|---|---|---|---|---|---|
| Density (g/mL) | ASTM D1218-12 [39] | 0.79 | 0.94 | 0.86 | 0.040 |
| Refractive index | | 1.27 | 1.50 | 1.46 | 0.029 |
| SimDis IBP(°C) | ASTM D2887-19ae2 [40] | 60 | 280 | 120 | 46.8 |
| SimDis T5(°C) | | 69 | 376 | 207 | 89.5 |
| SimDis T10(°C) | | 90 | 401 | 243 | 95.8 |
| SimDis T20(°C) | | 118 | 426 | 285 | 95.8 |
| SimDis T30(°C) | | 145 | 442 | 316 | 92.4 |
| SimDis T40(°C) | | 169 | 457 | 343 | 87.1 |
| SimDis T50(°C) | | 194 | 472 | 369 | 80.7 |
| SimDis T60(°C) | | 220 | 489 | 395 | 73.0 |
| SimDis T70(°C) | | 251 | 507 | 422 | 64.9 |
| SimDis T80(°C) | | 283 | 529 | 452 | 55.9 |
| SimDis T90(°C) | | 329 | 554 | 488 | 46.7 |
| SimDis T95(°C) | | 367 | 585 | 516 | 41.3 |
| SimDis FBP(°C) | | 472 | 661 | 581 | 27.8 |
| Conversion 370°C$^{+}$ | Calculated from SimDis | 3.4 | 96.0 | 38.7 | 28.20 |

*Table App 4.4. General statistical information of the cetane number estimated on 64 diesel samples obtained from the hydrocracking process.*

| | Méthod | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Cetane number** | IFPEN | 35.1 | 67.9 | 53.0 | 9.17 |
| **Density (g/mL)** | ASTM D1218-12 | 0.82 | 0.91 | 0.86 | 0.029 |
| **SimDis T5 (°C)** | ASTM D2887-19 | 112 | 253.7 | 208.5 | 27.6 |
| **SimDis T95 (°C)** | ASTM D2887-19 | 246 | 431 | 367 | 14.3 |

*Table App 4.5. Variable selection methods applied on the NIR data block*

| Method | Acronym | Parameters |
|---|---|---|
| Variable Importance in Projection [18] | VIP | Automatic |
| Selectivity Ratio [19] | SR | |
| Genetic Algorithm [21] | GA | Windowwidth = 50, mutation rate = 0.005, 30% initial terms, Convergence = 50% Algorithm = PLS (20LVs), 15 runs |
| Forward interval PLS [20] | iPLS (Forward & Backward) | Interval size [25,50,100,200] |
| recursive PLS [54] | rPLS | Max. interation = 500, Max. LVs = 20 |
| Covariance Selection [22] | CovSel | Features tuning [1-44] |

*Table App 4.6. Variable selection methods applied on the process variables data block*

| Method | Acronym | Parameters |
|---|---|---|
| Variable Importance in Projection | VIP | Automatic feature selection |
| Selectivity Ratio | SR | |
| Least Absolute Shrinkage and Selection Operator [33] | LASSO | Alpha tuning [0.01,0.07,0.05, 0.1, 1,2, 3, 5, 10] |
| Genetic Algorithm | GA | Windowwidth = 1, mutation rate = 0.005, 30% initial terms, Convergence = 50% Algorithm = MLR, 15 runs |
| Recursive Feature Elimination [34] | RFE | Features tuning [9-22] |
| eXtreme Gradient Boosting Feature Selection [55] | XGBoost_FS | Algorithm = gblinear, automatic feature selection |
| Sequential Feature Selection [35] | SFS (Forward & Backward) | Algorithm = MLR, automatic feature selection |
| Sequential Floating Feature Selection [36] | SFFS (Forward & Backward) | |
| Covariance Selection | CovSel | Features tuning [1-44] |

# Appendix 5: Publication 5. "Application of orthogonalization methods for robust diesel cetane number estimation from hydrocracking total effluent NIR spectra"

# Application of orthogonalization methods for robust diesel cetane number estimation from hydrocracking total effluent NIR spectra

J. Buendia-Garcia[a,c], M. Lacoue-Negre[a,c], J. Gornay[a], S. Mas-Garcia[b,c], R. Bendoula[b,c], J.M Roger[b,c]
[a] IFP Energies Nouvelles, Rond Point de l'échangeur de Solaize, France
[b] ITAP-INRAE, Institut Agro, University Montpellier, Montpellier, France
[c] ChemHouse Research Group, Montpellier, France

Corresponding Authors:
Jean-Michel Roger (jean-michel.roger@inrae.fr)

## Abstract

The accuracy of NIR models for predicting cetane number can be affected by external parameters related to spectrum acquisition. In this article, robust modelling to address this problem is investigated. This study evaluated the effectiveness of the external parameter orthogonalization (EPO) and dynamic orthogonal projection (DOP) methods in simultaneously correcting the impact of different external parameters affecting the performance of a NIR model for predicting diesel cetane number from spectra acquired on the hydrocracking (HCK) process total effluent. The impact of two types of optical instruments (probe and flow cell), two optical lengths (1 and 2 mm), four sample temperature levels (60°C – 90°C), and the effects caused by dynamic acquisition were analyzed using 444 spectra acquired on 129 samples. A reference partial least squares (PLS) model was developed using the spectra acquired at steady and constant conditions. Then a first model orthogonalization was conducted by applying the EPO method. As a result, the RMSEP was reduced by 76%. However, the effect caused by dynamic acquisition was slightly corrected. By applying a synergic orthogonalization using the DOP and EPO methods integrated, it was possible to obtain a robust model with RMSEP values (2.1) below the reproducibility of the reference method and reducing the bias caused by the external parameters by 99%. The robust modelling investigated in this study can be applied to estimating other diesel properties and characterizing other fuel products. In addition to the model robustness achieved, orthogonalization methods have a great advantage because no further processing or transformation of the new spectra is required, facilitating future model maintenance over time.

## Keywords

Orthogonalization, Near-Infrared (NIR), robustness, external parameters, hydrocracking total effluent, diesel fuel, cetane number.

# 1. Introduction

Near-infrared spectroscopy (NIR) is a fast and non-destructive analytical technique requiring minimal sample preparation that has been widely used in the last decades in the energy sector as an efficient alternative in the estimation of properties of crude oil [1], fuels [2–5], and biofuels [6–8] with errors close to the reproducibility of the standard reference methods.

The recent advances in the chemometrics domain have contributed to the recent growth in using NIR spectroscopy for developing more accurate predictive models. However, one of the main challenges encountered is to achieve the model robustness, defined as the ability to maintain a reliable performance under different application conditions [9, 10]. Generally, chemometric models are calibrated using databases with data acquired under stabilized and controlled laboratory conditions that adequately reflect the behavior of the studied variables. Hence, the model suitability is guaranteed if the information of new samples is obtained under repeatable conditions. Nevertheless, spectroscopic data, especially NIR, are very sensitive to any change in the acquisition conditions, and the model's predictive performance may be affected [11].

Although multiple acquisition conditions, also known as external parameters, can affect the spectrum quality and model performance, the most representatives can be classified into two groups. Parameters related to the experimental equipment used in the acquisition, such as the instrument type (probe, flow cell, reflectance, transmittance, optical length), and those related to the environment, such as humidity and temperature. Due to its high impact on spectra acquisition accuracy, temperature is the most studied influencing parameter [12, 13]. For instance, Hansen et al. [14] showed that molecular bonds vibration intensity depends on temperature, leading to changes in the spectrum according to temperature variation. Furthermore, some physicochemical properties of samples, such as viscosity and density, are temperature-dependent, and many changes in the sample due to temperature are not permanent and do not reflect the intrinsic nature of the sample [15–17]. Nevertheless, these changes can significantly affect spectrum acquisition. A third general parameter that can impact the model performance is the continuous spectral acquisition performed during real-time process monitoring. Even ensuring non-instrumental disturbances, the dynamic factors present during the spectra acquisition may lead to interferences, leading to random errors and deviations.

The spectral variability generated by an external parameter must be corrected, or at least minimized, to ensure a reliable description of the sample physicochemical behavior from the spectroscopic information extracted. In chemometrics, this issue is known as calibration transfer, or calibration adaptation, and is generally divided into four levels according to the problem to be corrected, bias, slope, dispersion and nonlinearities. Accordingly, Chauchard et al. [11] proposed a general methodology to determine the best strategy to correct the influence caused by an external parameter. In summary, when identifying an external parameter (G) with potential impact on the model performance, it must be determined whether the influence is significant or negligible. If it is a highly influential parameter, the next step is to determine if it can be controlled. Finally, in case of a negative outcome, it must be established whether the value (g) of G is known when using the model. If g is known, there are three strategies for correcting the influence of this parameter; a priori, a posteriori, and model correction. Otherwise, robust modelling must be conducted.

The a priori correction strategy, as its name indicates, is focused on the correction of the new spectrum

before its use. This correction is based on the difference between the existing spectra and the new spectrum affected by the G parameter. The most employed methods are the piecewise direct standardization (PDS) [18] and the spectral space transformation (SST) [19]. Concerning the model correction strategy, local modelling can be taken as an example. In this case, different regression models are generated with consolidated databases at different acquisition conditions defined by g. When a new sample is evaluated, its value is predicted from the sum of the $(\hat{y})$ obtained in each developed model, adjusted by a differential factor between the g used in the model development and the g of the new sample. Finally, the most frequent application of the a posteriori correction strategy is adjusting the value predicted by correcting the bias and slope, which are affected by the influence of G.

In the oil & gas industry, some studies related to the correction of external parameters affecting the model performance for predicting fuel properties can be found. In their study, Pereira et al. [20] compared five calibration transfer methods (DS, PDS, orthogonal signal correction (OSC), reverse standardization (RS), and piecewise reverse standardization (PRS)), to correct the impact of instrument change. They searched to estimate two gasoline properties (naphthenes content and research octane number (RON)) from NIR spectra acquired in 3 different NIR spectrometers. In addition to the a priori correction strategy evaluated, they also analyzed the slope/bias correction and model update strategies. The best results were obtained by using RS.

Another study that used the RS as a calibration transfer method for correcting the impact generated by instrument change was developed by da Silva et al. [21]. They corrected the impact in predicting five fuel quality parameters (density and the simulated distillation (SimDis) temperatures (T) to recover the 10%, 50%, 90% of the sample and the final boiling point (FBP)). Moreover, in a complementary work to the developed by Cooper et al., [22] Abdelkader et al., [23] compared the PDS method and the slope/bias adjustment strategy to correct the influence of this external parameter in estimating 13 jet fuel properties. In their study, they showed that the two strategies present similar results in 6 (SimDis T10, T20, T50, flash point, freezing point, hydrogen content) of the 13 properties, while in the remaining 7 (API gravity, cetane index, saturates, aromatics, density, SimDis T90, viscosity) the effect of the external parameter is better corrected by the slope and bias adjustment. Complementing the studies conducted to correct the instrument change impact on model performance, it can be highlighted the work of Rodrigues et al.[24] They compared the DS, PDS and the orthogonal projections in latent structures (OPLS) methods, being the latter the one that presented better results in correcting the influence of this external parameter in the prediction of oil density.

Regarding temperature impact correction, most of the studies are developed in the food and agricultural sector [11, 25, 26], while in the oil & gas area, the literature is very limited. Although not directly related to this industry, the study done by Haroon et al. [27] can give a glimpse of the different methods for correcting the temperature influence on NIR model performance for property prediction in liquid samples. They evaluated the a priori correction strategy using the generalized least squares weighting (GLSW) and DS

methods. They also evaluated the model correction strategy using two different approaches (temperature as dependent and independent variable). This study concluded that the method that best corrected the effect of temperature on predicting micellar liquid viscosity was the GLSW. Related to the fuels field, Baird et al. [28] employed the a posteriori correction strategy to correct the temperature effect on gasoline and diesel density prediction. They first developed a model using the support vector regression (SVR) method from a calibration data set acquired at a defined and constant temperature (20°C). Then using the same SVR method, they developed a second model to predict the slope coefficient used in correcting the value predicted by the first model.

Concerning the correction of the dynamic spectra acquisition impact, Macho et al. [29] developed a model to monitor in real-time the ethylene content in polypropylene polymers which, after an optimal performance window of 55 days, was affected by an instrumental drift caused by a change in the wavelength position. The impact generated by this parameter was corrected using a model slope/bias correction.

A strategy less employed to correct the impact of external parameters on the model performance is the robust modelling strategy. Compared to the strategies described previously, robust modelling is an alternative having higher efficiency in solving the problem of the constant and sometimes unexpected variability that may occur in the acquisition of a spectrum. This is because its basic principle is the adaptation (transformation) of the domain of the calibration data set, which implies a non-need to transform, correct or adapt the spectral information of the new sample evaluated. In this strategy, it can be found mainly the orthogonalization methods, such as the external parameter orthogonalization (EPO) [30] and the dynamic orthogonal projection (DOP) [31].

From the reviewed literature concerning the correction of external parameters impact on NIR model performance in the oil & gas industry, it was found that generally the correction of these parameters is investigated separately. In other words, one parameter is corrected at a time. Another finding that can be highlighted is that the reported studies use the first three strategies (a priori, a posteriori, and model correction) to correct the calibration adaptation problem. The only study found in the oil & gas industry that corrects multiple parameters at the same time (instrumental disturbance and sample temperature) using robust modelling is the one reported by Amat-Tosello et al [32]. In their study they used the EPO method to obtain a robust predictive model to estimate the research octane number (RON) and the motor octane number (MON) of gasoline from the NIR spectra acquired in four different instruments. Their results showed the advantage of the EPO method in not needing a transfer function to transform the new spectra.

Considering the importance of correcting the external parameters influence to ensure the robustness of prediction models, the study presented in this article evaluated the potential of two orthogonalization methods (EPO and DOP) to simultaneously correct the impact of three external parameters: (i) instrumental

disturbances (instrument type (probe and flow cell) and optical length (1 and 2 mm)), (ii) sample temperature, and (iii) influence of dynamic spectra acquisition on the diesel cetane number estimation. The evaluation was carried out in 3 steps: the generation of the database containing the spectra obtained at different acquisition conditions, the development of the baseline model for assessing the impact of the external parameters studied, and the development of the robust model by applying the orthogonalization methods described before. As a result, in this study was developed a robust model for predicting the property investigated from NIR spectra acquired on the total effluent of the HCK process. The context of predicting the diesel cetane number from NIR spectra of the total effluent can be found in [33].

# 2. Materials and methods

In this study, two types of samples were considered; the total effluent obtained from the HCK process for NIR spectra acquisition and the diesel recovered from the distillation of the total effluent on which the cetane number was measured [33]. This section details the experiments conducted to obtain each type of sample, the laboratory analyses carried out on each of them, and the methodology for developing the models.

## 2.1 Total effluent samples

Generally, the total effluent samples are obtained when heavy crude oil residues, mostly vacuum gas oils (VGO's), are processed in the HCK process reactors. In this study, 129 total effluent samples were obtained by processing 29 different feedstocks in the HCK pilot plants at the IFPEN in Solaize, France, under different operating conditions ensuring that the diversity of the total effluent physicochemical properties was representative of the different HCK process scenarios [33].

The NIR spectra used in this study were acquired on the 129 total effluent samples obtained and under two general acquisition categories (steady and dynamic). The detailed acquisition conditions and the number of spectra acquired are summarized in Table 2.

*Table 2. NIR spectra acquisition conditions*

| Instrument | Optical lenght (mm) | Sample temperature (°C) | Condition | Spectra aquired | Use |
|---|---|---|---|---|---|
| Probe | 2 | 60 | Static | 98 | Baseline modelling |
| | | 60 | | 27 | Orthogonalization |
| | | 70 | | 27 | |
| | | 80 | | 27 | |
| | | 90 | | 27 | |
| | 1 | 60 | | 27 | |
| Flowcell | 1 | 60-90 | Dynamic | 211 | |

Before to describe each acquisition category, it is important to mention that all spectra were acquired using a NIRS XDS Process Analyzer (Metrohm, Villebon - France) spectrometer, recording wavelengths from 800 - 2200 nm with a resolution of 0.5 nm. 32 scans were acquired on the sample and then averaged to produce

the final spectrum. The software used with the spectrometer was VISION (Metrohm, Villebon - France). Moreover, before any spectra acquisition (in steady or dynamic conditions), the total effluent samples were heated in a water bath at 60°C in a closed flask for one hour and then manually shaken to ensure their liquid state and homogeneity. In the steady acquisition of spectra at temperatures other than 60°C, the sample was heated for thirty additional minutes at the desired temperature.

### Spectra acquisition at steady conditions

The NIR spectra acquired at steady conditions were divided into two groups. The first corresponds to the spectra used for calibration of the reference model. A total of 98 total effluent samples were gathered in this group, and their spectra were acquired at a constant sample temperature of 60°C and using a reflectance Falcata Lab6 probe (Hellma GmbH & Co. KG, Müllheim - Germany) with an optical length of 2 mm. The second group corresponds to the spectra used in the model orthogonalization. In this group, 135 NIR spectra acquired on 27 total effluent samples at different conditions were consolidated. Firstly, using the same probe as described before, 108 spectra were acquired by varying the sample temperature between 60°C and 90°C at a ΔT of 10°C. Then, using a 1 mm optical length in the Falcata probe and a sample temperature of 60°C, the remaining 27 spectra were acquired.

### Spectra acquisition at dynamic conditions

For the dynamic acquisition of the NIR spectra, a closed circulation loop with a 1/8" OD tubing was used. A 305 HPLC pump (Gilson, Villiers le bel, France) was employed to set the sample flow rate at 10 ml/min. A water bath regulated the sample temperature with an integrated control system. A heating ribbon was installed on the tubing lines and instruments and then coated with glass fiber as an insulating material to avoid significant heat losses during the sample recirculation. A transmission NIR Flow cell 1/4" OD tube (Metrohm, Villebon - France) with an optical length of 1 mm was employed for the spectra acquisition.

The samples were heated at 60°C in the water bath before the spectra acquisition and recirculated for 10 minutes. Next, the NIR spectra were acquired every 30 seconds during 30 minutes, increasing the sample temperature by 10°C every 10 minutes. This procedure was applied on four new samples of total effluent, i.e., different from the 125 samples analyzed in the steady acquisition.

## 2.2 Diesel samples

The diesel samples used in this study were recovered after distillation of each of the total effluent samples using the ASTM D2892-20 standard [34]. The diverse quality of the recovered diesel samples was assured. Table 3 presents four statistical parameters calculated on the diesel density [35] and the simulated distillation temperatures range to obtain both 5% and 95% of sample distillate (Simulated Distillation T5 and T95) [36]. Using an internal method developed in the IFPEN validated against the ASTM D613 standard method,[37] the cetane number was estimated on the diesel samples. The general statistical information of this property is also shown in Table 3.

*Table 3. General statistical information of the cetane number estimated on 98 diesel samples obtained from the hydrocracking process*

| | Méthod | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Cetane number** | PackPIR | 30.3 | 69.5 | 51.6 | 11.1 |
| **Density (gr/ml)** | ASTM D1218-12 | 0.8100 | 0.9114 | 0.8604 | 0.0312 |
| **SimDis T5 (°C)** | ASTM D2887-19 | 212.6 | 257.6 | 245 | 9.1 |
| **SimDis T95 (°C)** | ASTM D2887-19 | 245.8 | 430.9 | 366.6 | 15.3 |

## 2.3 Modelling

This study employed the spectral region between 1110 and 2200 nm taking into account that this region of the total effluent NIR spectrum provides the most descriptive information to predict the diesel properties [33, 38, 39].

In this article it was evaluated the effectiveness of two orthogonalization methods (EPO and DOP) to correct simultaneously the impact of three external parameters on the model performance for diesel cetane number estimation. Namely, the sample temperature, the instrument perturbance (type and optical lenght), and the dynamic acquisition conditions. Therefore, the work was developed implementing seven modelling steps. Table 4 summarizes the information used in the development of the study.

*Table 4. Summary of NIR spectra employed in each modelling step*

<table>
<tr><th rowspan="2">#Spectra</th><th rowspan="2">Acquisition conditions</th><th rowspan="2">Data set</th><th colspan="2">Cetane number</th><th rowspan="2">Modelling step</th></tr>
<tr><th>Min</th><th>Max</th></tr>
<tr><td>67</td><td>Sample temperature = 60°C<br>Instrument = Probe<br>Optical length = 2 mm</td><td rowspan="2">Calibration</td><td rowspan="2">30.3</td><td rowspan="2">69.5</td><td>Baseline model</td></tr>
<tr><td>91</td><td rowspan="2">Sample temperature = 60°C - 90°C<br>Instrument = Probe & Flow cell<br>Optical length = 1 & 2 mm</td><td>Orthogonalization</td></tr>
<tr><td>286</td><td>Test</td><td>35.4</td><td>69.3</td><td>Tets reference and orthogonalized models</td></tr>
</table>

Step 1: Considering the information described in section 2.1 and summarized in Table 2, the first modelling step was completed by generating three sets of data corresponding to the information of the independent variables (NIR spectra). The Mx1 matrix containing the spectra acquired at steady and constant temperature conditions, the Mx2 matrix containing the spectra acquired at steady conditions but varying the temperature and the probe optical length, and the Mx3 matrix corresponding to the spectra acquired under dynamic conditions. For each NIR matrix generated, its respective dependent variable matrix (My) was generated.

Step 2: The second implemented step involved the preprocessing of the Mx matrices through the combination of the Standard Normal Variate (SNV) [40] and the second derivative of Savitzky-Golay with a third polynomial order and a 23 window-point (SavGol[23,3,2]) [41].

Step 3: The third step was the definition of the different sub-datasets employed in the study as follows. The Mx1 was split into the calibration (Mx1C, 70% samples) and the test (Mx1T, 30% samples) sub-datasets using the Kennard-Stone (KS) algorithm [42]. In turn, the Mx2 was sorted in ascending order regarding the

corresponding My values. Then taking 1 out of 3 samples, the Mx2 with its associated My was divided into four sub-datasets: (i) Mx2DC (NIR spectra acquired at steady conditions using a constant temperature of 60°C and an optic length of 2mm), (ii) Mx2DV (NIR spectra acquired on the same samples of Mx2DC but varying sample temperature and the optic length), (iii) Mx2A and (iv) Mx2T (NIR spectra acquired on the remaining samples varying sample temperature and the optic length). Finally, the Mx2T and the test dataset defined in the PLS reference model were concatenated to test the developed models.

Step 4: A reference partial least squares (PLS) model was developed based on the calibration sub-dataset defined from Mx1 matrix (Mx1C). The number of latent variables (LV's) retained in the developed PLS model was defined using the RMSECV as the Figure of merit [33].

Step 5: In the fifth step a first orthogonalization of the reference model was implemented utilizing the method EPO [30] to achieve the robust prediction of the diesel cetane number. To do so, the detrimental matrix D representing the influence spectra was calculated using the preprocessed Mx2DC and Mx2DV matrices (D = Mx2DC - Mx2DV). Next, the optimal combination of the number of EPO components (EPOC) and the number of LV's to be used were determined evaluating different orthogonalized PLS models obtained from the calibration dataset (MxC_EPO) composed of the matrices Mx1C, Mx2DC, Mx2DV using the SEP as a figure of merit calculated from the Mx2A matrix. Finally, the MxC_EPO matrix was orthogonalized using the defined EPOC's, and a PLS model was calibrated, retaining the established LV's. In summary, the matrices Mx2DC and Mx2DV were used to calculate the detrimental matrix D employed in the orthogonalization of the model, while the Mx2A matrix was considered to adjust the number of the EPO method components used and the LV's of the orthogonalized model retained.

Step 6: A second orthogonalization modelling focused on synergic applying the DOP and EPO methods [31] to correct any disturbances generated by the dynamic acquisition of the spectra that were not captured in the initial orthogonalization of the model was implemented. In this step, the Mx3 and its corresponding My were employed. First, the DOP method was applied on a sample spectrum analyzed in dynamic conditions to obtain the corresponding "ideal" spectrum. Then, the "ideal" spectrum calculated with the DOP method is included in the Mx2DC sub-dataset, and the spectrum acquired under dynamic conditions is integrated into the Mx2DV sub-dataset. With these updated sets of data, the steps previously described to apply the EPO method (see step 6) were implemented once again. Knowing the measured value of the variable (y) of the analyzed samples, one spectrum corresponding to a sample within the worst dynamic prediction group was randomly selected to apply the DOP method (spectrum 30 sample 4).

Step 7: The last step was the performance evaluation of the different models developed. Using the test sub-dataset generated in step 2, all models were tested. This sub-dataset contains the spectra acquired a steady and dynamic conditions. The figures of merit used to evaluate the model performance was the standard error

of prediction (SEP), the bias, and the root mean squared error of prediction (RMSEP). The reduced Q residual and the Hotelling $T^2$ analysis were employed complementary to evaluate the suitability of the models for predicting the diesel cetane number from NIR spectra acquired at different conditions.

The models and variable selection analysis were performed using the PLS_Toolbox V.8.8 (Eigenvector Research Inc. Wenatchee, WA, USA), MATLAB V.2019b (The MathWorks, Inc., Natick, MA, USA).

# 3. Results and discussion

The results and their discussion were divided into two main parts: developing the reference model and applying the orthogonalization methods. Table 5 presents the statistical information calculated for the final comparison of the developed models' performance (reference vs. orthogonalized).

## 3.1 Baseline model calibration

A PLS model with 9 LV's was calibrated from the 67 spectra acquired at steady and constant conditions (see Table 4). This model captures 99% of the X and Y matrices explained variance, presenting values of root mean squared error of calibration (RMSEC = 1.3) and cross-validation (RMSECV = 2.2) below the reproducibility limits of the reference method (±3.6). Furthermore, the respective squared correlation coefficients ($r^2C$ & $r^2C$) are greater than 0.96. For determining the impact of external parameters on the diesel cetane number estimation, the model performance was evaluated using the 286 spectra described in Table 4 (76 acquired at steady conditions and 210 at dynamic conditions).

Table 5 reports that the RMSEP of the model when using the entire test data set is 24.1 with a bias of -21.7 (SEP = 10.5). These values are mostly due to the estimation of cetane number in samples whose NIR spectra were acquired at varying conditions. Figure 3a shows that 99% of these samples were predicted outside the reproducibility limits of the reference method, i.e., 284 out of 286 samples. The samples predicted within the defined limits correspond to two of the nine samples having spectra acquired using the probe with a path length of 1mm. The samples with the highest error in the property estimation are those whose NIR spectra were acquired under dynamic conditions. In contrast, all but one of the samples whose NIR spectra had been acquired at the same conditions as those employed in the model calibration data set (sample temperature = 60°C, probe, and path length = 2mm) were predicted within these limits; representing a 98% effectiveness in cetane number estimation at these conditions.

When performing the graphical analysis of the Qresidual and Hotelling $T^2$ statistical tests applied on the test samples using the 9LV's of the developed PLS model, the significant impact of the external parameters on the quality of the spectra can be validated. Figure 4a shows that 98% of the samples with NIR acquired at variable and dynamic conditions exceed the threshold of the two tests. This Figure also validates that dynamic conditions greatly impact the spectra quality.

Typically, samples identified above the threshold of these two tests would be categorized as outliers and should be discarded or, since the atypical behavior is due to the acquisition conditions and not properly to an error in the measurement or recording of the information, the spectra of these samples should be re-acquired at the stable and steady conditions used in the calibration set. Another possible solution is the development of transfer functions to transform the acquired spectra at varying conditions so that they can be used with confidence in the model. The main shortcomings of these options are the cost of experimentation and the continuous reprocessing of the information, especially with the variability that may occur in acquiring new spectra. Additionally, combining different external parameters makes this task even more complex. Therefore, a more efficient solution must be applied.

## 3.2 Model orthogonalization

A first orthogonalization for correcting the impact of external parameters on the developed PLS model was performed by applying the EPO method. With a Matlab script adapted from Roger et al., [30] this first orthogonalization was implemented as described in section 2.3 step 5.

Figure 1 shows the SEP's obtained from the different combinations of EPOC's and LV's, evidencing a region with SEP's below 2.0 between EPOC's 9 & 12 and LV's 7 & 9. The combination with the lowest SEP (1.2) was 12 EPOc's and 7 LV's.

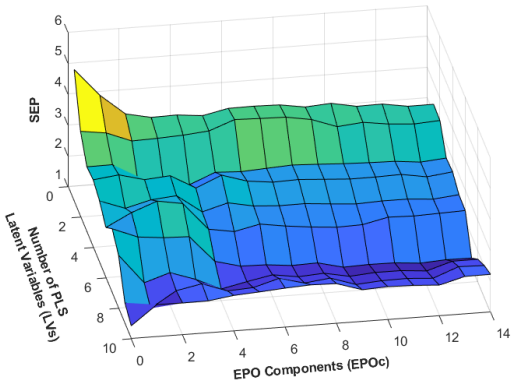



*Figure 1. Standard error of prediction on Mx2A as a function of the number of PLS latent variables and EPO components*

The orthogonalized PLS model with the EPO method (PLS_EPO) presented a slightly higher RMSEC (1.6) than the baseline PLS model. This model "deterioration" is evidenced when, compared with the reference PLS model, the sample with the lowest cetane number is predicted outside the reproducibility limits, even when its spectrum has been acquired at steady and constant conditions (see Figure 3b). On the contrary, the RMSECV remained nearly the same (2.0). The performance of the PLS_EPO model was evaluated with the same 286 spectra previously used in testing the reference PLS model. In Figure 3b, it is observed that the prediction of the samples whose NIR spectra were acquired at varying conditions is improved. Out of 9 samples acquired at 60°C and using an optical length of 1mm, 8 are predicted within the reproducibility limits of the reference method. The correction of the impact of sample temperature is also evident. Regardless of

the value employed on this parameter, the cetane values are satisfactorily predicted, except for the sample with the lowest cetane value. Two of the four samples analyzed under dynamic conditions are predicted close to the reproducibility limits, while the other two are still predicted outside these limits. An important fact to highlight is that despite the "model deterioration," the effectiveness in predicting the samples analyzed at 60°C and optical length of 2 mm is barely affected.

Table 5 summarizes the statistical information of this model where it can be observed that as a result of the correction achieved with the model orthogonalization, the RMSEP (5.8) and the bias (-4.18, SEP = 4.0) are significantly reduced. At the same time, the squared correlation coefficient of prediction is improved ($r^2P$, 0.698 vs. 0.019). However, the RMSEP value is still higher than the reproducibility limit of the reference method. This value is mainly due to the samples analyzed in dynamic conditions predicted outside these limits. Figure 4b shows that these samples are still having a significant impact on model performance. Although the conditions used to acquire the spectra under dynamic conditions are similar to those used under steady conditions (sample temperature = 60°C - 90°C, optical length = 1 & 2mm), it should be noted that the instrument is different. While a reflectance probe was used in the steady acquisitions, a transmittance flow-cell was used in the dynamic acquisitions. The results show that this instrument change also significantly affects the quality of the spectra, impacting the model performance.

The impact caused by the instrument type could be corrected in the orthogonalization of the model. Nevertheless, for this case, the EPO method cannot be applied directly to the calibration dataset since there is no reference spectrum of these samples, i.e., a spectrum acquired at 60°C and optical length of 2 mm, impeding an updated D matrix calculation. To mitigate this problem, the DOP method plays an essential part.

To correct the impact of the instrument type, and any other effects that acquisition under dynamic conditions may be having on model performance, an integrated orthogonalization was performed using the EPO and DOP methods in a complementary approach (see section 2.3 step 6).

Figure 2 shows the SEP's obtained from the EPOc's and LV's combinations evaluated in this integrated orthogonalization. In this second analysis, the region with SEP's below 2.0 is observed between EPOc's 9 & 14 and LV's 8 & 10. The combination with the lowest SEP's (1.3) was 12 EPOc's and 8 LV's.

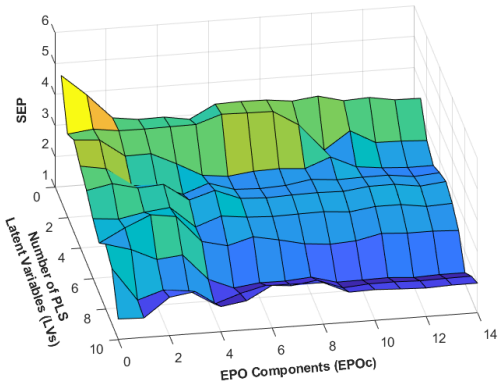



*Figure 2. Standard error of prediction on Mx2A as a function of the number of PLS latent variables and EPO components when DOP is applied*

The new orthogonalized PLS model using the EPO and DOP methods combined (PLS_EPO_DOP) presents an RMSEC (1.6) and an RMSECV (2.0) equal to those shown by the PLS_EPO model. These values validate the consistency in the application of the orthogonalization methods. Regarding the RMSEP (2.1) and bias (-0.285, SEP = 2.1), the PLS_EPO_DOP model reduces them to values close to and even lower than the reproducibility of the reference method (±3.6). The parity plot shown in Figure 3c shows how, with some exceptions, the impact of all the parameters evaluated in this study is corrected, resulting in a robust model. This is also reflected in the significant improvement of the $r^2P$ (0.917).

Figure 4c also shows how the impact of the different parameters is compensated, and the Qresidual vs. Hotelling $T^2$ analysis can be performed objectively and reliably to determine the samples with atypical behavior. For example, the five estimated values that significantly exceed the threshold of the two tests correspond accurately to the sample with the lowest cetane number, which was predicted slightly outside the reproducibility limits. This sample corresponds to a set of four samples analyzed during an HCK test, being the only one with a poor prediction. The difference between these four samples is the temperature used in the HCK reactor. It is normally expected that the higher the reaction temperature, the higher the cetane number of the diesel. While this trend is observed in the three correctly estimated samples, the predicted sample outside the limits slightly disrupts this trend. Therefore, the poor prediction could be attributed to the disrupted trend. Nonetheless, as discussed before, the poor prediction could be rather caused for the " model deterioration" when the orthogonalization is applied.

*Table 5. Statistical parameters summary for cetane number models comparison*

| | PLS | PLS_EPO | PLS_EPO_DOP |
|---|---|---|---|
| **#EPO$_C$** | - | 12 | 12 |
| **#LV's** | 9 | 7 | 8 |
| **RMSEC** | 1.3 | 1.6 | 1.6 |
| **BiasC** | 0.00 | 0.04 | -0.01 |
| **r²C** | 0.986 | 0.981 | 0.981 |
| **RMSECV** | 2.1 | 2.0 | 2.0 |
| **BiasCV** | -0.05 | -0.01 | -0.01 |
| **r²CV** | 0.962 | 0.970 | 0.969 |
| **RMSEP** | 24.1 | 5.8 | 2.1 |
| **BiasP** | -21.70 | -4.18 | -0.28 |
| **r²P** | 0.019 | 0.698 | 0.917 |
| **SEP** | 10.5 | 4.0 | 2.1 |

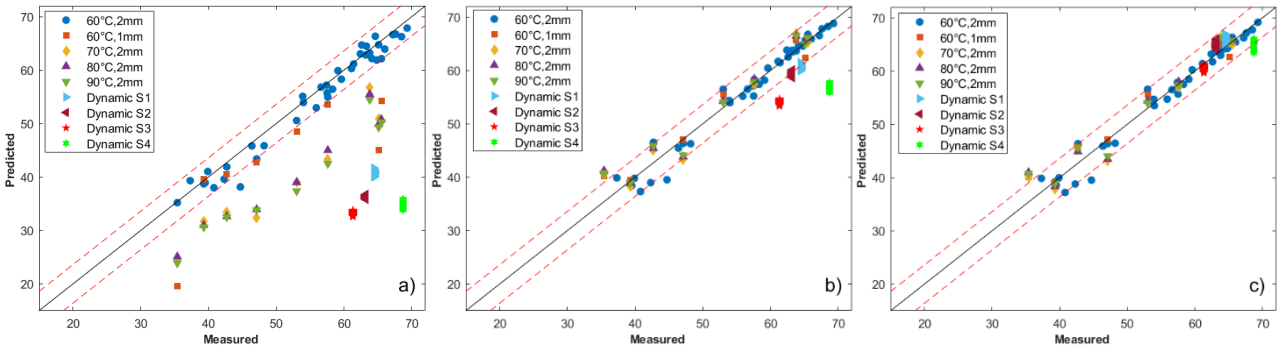



*Figure 3. Parity chart of cetane number estimations. a) PLS model, b) PLS_EPO model, c) PLS_EPO_DOP model*

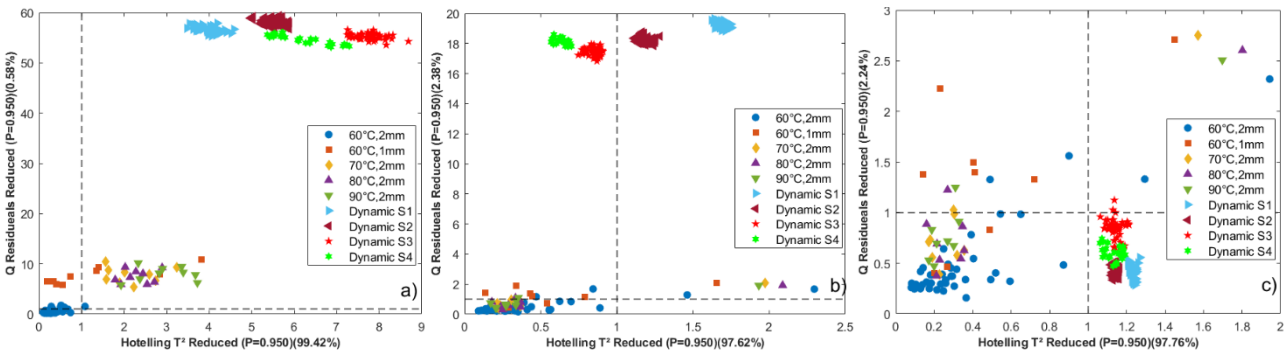



*Figure 4. Reduced Q residual and Hotelling T² analysis. a) 9 LVs PLS model, b) 7 LVs PLS_EPO model, c) 8LVs PLS_EPO_DOP model*

# Conclusions

This paper evaluated the efficiency of the EPO and DOP orthogonalization methods to compensate the impact of three external parameters on the model performance for diesel cetane number prediction from NIR spectra of HCK total effluent. Namely, sample temperature, instrument optical length, instrument type, and acquisition conditions (steady and dynamic).

The two external parameters of greatest influence on the spectra quality and therefore on the model performance are the sample temperature and the acquisition under dynamic conditions, the latter having the greatest impact. On the contrary, the influence of the probe path length is of the least consequence.

The EPO orthogonalization method significantly corrects the impact of the external parameters studied in this article. The influence of sample temperature and path length is the best corrected. Although these two parameters are immersed in the correction of the dynamic acquisition conditions impact, the type of instrument and other effects inherent to the dynamics, such as flow rate, are not fully corrected. This limitation is overcome by integrating the DOP method.

The orthogonalized prediction model shows a slight deterioration compared to the baseline model in predicting samples analyzed at steady-state conditions. However, this deterioration is negligible compared to the robustness achieved in the model.

The integrated application of the orthogonalization methods showed that regardless of the acquisition conditions, stable or variable, steady or dynamic, and even with different types of instrument, the predictions of the diesel cetane number are reliable over the whole range of estimation evaluated. In addition to the model robustness achieved, a great advantage of using orthogonalization methods is that no further processing or transformation of the new spectra is required, facilitating the maintenance of the models over time.

# Acknowledgments

The authors would like to thank IFP Energies Nouvelles for providing the total effluent samples from the HCK process reactors, the facilities for the distillation to obtain the diesel samples, and the facilities for spectra acquisition and data analysis. Special thanks to Sebastien Giroud its valuable help in constructing the loop used for the dynamic analysis. Thanks also to Axel One Analysis for providing the probe for the NIR spectra acquisition.

# CRediT authorship contribution statement

**J. Buendia-Garcia**: Conceptualization, Data curation, Writing - original draft. **J. Gornay:** Conceptualization, Writing - original draft. **M. Lacoue-Negre:** Conceptualization, Writing - original draft. **S. Mas-Garcia:** Writing - original draft. **R. Bendoula:** Writing - original draft, **J.M Roger:** Conceptualization, Writing - original draft

# Declaration of conflicting interests

The Author(s) declare(s) that there is no conflict of interest

# REFERENCES

[1] M. K. Moro, F. D. dos Santos, G. S. Folli, W. Romão, P. R. Filgueiras, Fuel 2021, 303.
[2] N. Zanier-Szydlowski, A. Quignard, F. Baco, H. Biguerd, L. Carpot, F. Wahl, Oil & gas science and technology - rev IFP 1999, 54, 463–472.
[3] J. Li, X. Chu, Energy Fuels 2018, 32, 12013–12020.
[4] I. Hradecká, R. Velvarská, K. Dlasková Jaklová, A. Vráblík, Infrared Physics & Technology 2021.
[5] M. Gómez-Carracedo, J. Andrade, M. Calviño, E. Fernández, D. Prada, S. Muniategui, Fuel 2003, 82, 1211–1218.
[6] J. Skvaril, K. Kyprianidis, A. Avelin, M. Odlare, E. Dahlquist, Energy Procedia 2017, 105, 1309–1317.
[7] E. Wikberg, S. Heikkilä, K. Sirviö, P. Välisuo, S. Niemi, A. Niemi, Fuels 2021, 2, 179–193.

[8] M. Pilar Dorado, S. Pinzi, A. de Haro, R. Font, J. Garcia-Olmo, Fuel 2011, 90, 2321–2325.
[9] M. Zeaiter, J.-M. Roger, V. Bellon-Maurel, D. N. Rutledge, TrAC Trends in Analytical Chemistry 2004, 23, 157–170.
[10] M. Zeaiter, J.-M. Roger, V. Bellon-Maurel, TrAC Trends in Analytical Chemistry 2005, 24, 437–445.
[11] F. Chauchard, J.M. Roger and V. Bellon-Maurel, Journal of Near Infrared Spectroscopy 2004, 12, 199–205.
[12] Florian Wülfert,†, Wim Th. Kok,† and, and Age K. Smilde*,†, Analytical Chemistry 1998, 70, 1761–1767.
[13] Hideyuki Abe, Chie Iyo, and Sumio Kawano, Journal of Near Infrared Spectroscopy 2000, 8, 209–213.
[14] W.G. Hansen, S.C.C. Wiedemann, M. Snieder, and V.A.L. Wortel, Journal of Near Infrared Spectroscopy 2000, 8, 125–132.
[15] Jasem M. Al-Besharah/Saed A. Akashah/Clive J. Mumford, Industrial & Engineering Chemistry 1989, 213–221.
[16] P. Luo, Y. Gu, Fuel 2007, 86, 1069–1078.
[17] R. Payri, F. J. Salvador, J. Gimeno, G. Bracho, Fuel 2011, 90, 1172–1180.
[18] Y. Wang, D. J. Veltkamp, B. R. Kowalski, Anal. Chem. 2002, 63, 2750–2756.
[19] W. Du, Z.-P. Chen, L.-J. Zhong, S.-X. Wang, R.-Q. Yu, A. Nordon, D. Littlejohn, M. Holden, Analytica Chimica Acta 2011, 690, 64–70.
[20] C. F. Pereira, M. F. Pimentel, R. K. H. Galvão, F. A. Honorato, L. Stragevitch, M. N. Martins, Analytica Chimica Acta 2008, 611, 41–47.
[21] N. C. da Silva, C. J. Cavalcanti, F. A. Honorato, J. M. Amigo, M. F. Pimentel, Analytica Chimica Acta 2017, 954, 32–42.
[22] J. B. Cooper, C. M. Larkin, M. F. Abdelkader, Journal of Near Infrared Spectroscopy 2011, 19, 139–150.
[23] M. F. Abdelkader, J. B. Cooper, C. M. Larkin, Chemometrics and Intelligent Laboratory Systems 2012, 110, 64–73.
[24] R. R. Rodrigues, J. T. Rocha, L. M. S. Oliveira, J. C. M. Dias, E. I. Müller, E. V. Castro, P. R. Filgueiras, Chemometrics and Intelligent Laboratory Systems 2017, 166, 7–13.
[25] H. Kaur, R. Künnemeyer, A. McGlone, Molecules (Basel, Switzerland) 2022, 27.
[26] X. Sun, P. Subedi, K. B. Walsh, Postharvest Biology and Technology 2020, 162, 111117.
[27] K. Haroon, A. Arafeh, T. Rodgers, Ć. Mendoza, M. Baker, P. Martin, J. Chemometrics 2020, 34.
[28] Z. S. Baird, V. Oja, Chemometrics and Intelligent Laboratory Systems 2016, 158, 41–47.
[29] S. Macho, M. S. Larrechi, TrAC Trends in Analytical Chemistry 2002, 21, 799–806.
[30] J.-M. Roger, F. Chauchard, V. Bellon-Maurel, Chemometrics and Intelligent Laboratory Systems 2003, 66, 191–204.
[31] M. Zeaiter, J. M. Roger, V. Bellon-Maurel, Chemometrics intelligent laboratory systems 2005.
[32] S. Amat-Tosello, N. Dupuy, J. Kister, Analytica Chimica Acta 2009, 642, 6–11.
[33] J. Buendia Garcia, M. Lacoue-Negre, J. Gornay, S. Mas Garcia, R. Bendoula, J. M. Roger, Submitted to Fuel 2022.
[34] ASTM D 2892-20, Standard Test Method for Distillation of Crude Petroleum (15-Theoretical Plate Column), 2020, ASTM International, West Conshohocken, PA.
[35] ASTM D1218 - 12, Standard Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids, can be found under https://www.astm.org/Standards/D1218.htm.
[36] ASTM D2887 - 19ae1, Standard Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography, can be found under https://www.astm.org/Standards/D2887.htm.
[37] ASTM D613-01, Test Method for Cetane Number of Diesel Fuel Oil, 2001, ASTM International, West Conshohocken, PA.
[38] E. D. Yalvac, M. B. Seasholtz, S. R. Crouch, Appl. Spectrosc., AS 1997, 51, 1303–1310.
[39] J. J. Kelly, J. B. Callis, Anal. Chem. 1990, 62, 1444–1451.
[40] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Appl Spectrosc 1989, 43, 772–777.
[41] Abraham. Savitzky/M. J. E. Golay, Analytical Chemistry 1964, 36.
[42] R. W. Kennard, L. A. Stone, Technometrics 1969, 11, 137–148.

# Appendix 6: Detailed modelling results

*Table 1 Appendix 6. Description of the models*

| Model | Description | Acronym |
|---|---|---|
| 1 | NIR model | Mod_1 |
| 2 | NIR model with EPO correcttion | Mod_2 |
| 3 | NMR model | Mod_3 |
| 4 | NIR + NMR data fusion model | Mod_4 |
| 5 | NIR + PV_TE* data fusion model | Mod_5 |
| 6 | NIR + PV_NTE** data fusion model | Mod_6 |
| 7 | NIR + NMR + PV_TE data fusion model | Mod_7 |
| 8 | NIR + NMR + PV_NTE data fusion model | Mod_8 |

**PV_TE = Process Variables including Total Effluente properties*
**PV_NTE = Process Variables not including Total Effluente properties*

*Table 2 Appendix 6. Diesel cetane number modelling results*

| Variable predicted | **Cetane Number** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±3.6 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Regression Method | PLS | | | | PLS + MLR | | | |
| Pre-processing scheme | PP1 | | PP2 | PP3 | PP4 | | PP5 | |
| Data fusion level | - | - | - | Mid | High | High | High | High |
| EPO components | - | 12 | - | - | - | - | - | - |
| Calibration Data Points | 67 | 139 | 67 | 67 | 51 | 51 | 51 | 51 |
| Test Data Points | 31 | 31 | 31 | 31 | 23 | 23 | 23 | 23 |
| Total Data Points | 98 | 170 | 98 | 98 | 74 | 74 | 74 | 74 |
| NIR Variables | 2180 | 2180 | - | 7 LVs scores | ŷ 7LVs PLS | ŷ 10LVs PLS | ŷ 12LVs PLS | ŷ 12LVs PLS |
| NMR Variables | - | - | 13926 | 7 LVs scores | - | - | ŷ 8LVs PLS | ŷ 8LVs PLS |
| Process Variables | - | - | - | - | ŷ MLR | ŷ MLR | ŷ MLR | ŷ MLR |
| Latent Variables | 9 | 7 | 7 | 11 | - | - | - | - |
| Calibration $R^2$ | 0.986 | 0.982 | 0.982 | 0.992 | 0.995 | 0.991 | 0.996 | 0.992 |
| Cross Validation $R^2$ | 0.959 | 0.973 | 0.949 | 0.986 | 0.995 | 0.989 | 0.993 | 0.988 |
| Prediction $R^2$ | 0.966 | 0.982 | 0.973 | 0.984 | 0.988 | 0.975 | 0.993 | 0.987 |
| RMSEC | 1.3 | 1.6 | 1.3 | 0.9 | 0.6 | 0.9 | 0.6 | 0.8 |
| RMSECV | 2.2 | 1.9 | 2.2 | 1.1 | 0.7 | 1.0 | 0.7 | 1.0 |
| RMSEP | 2.0 | 1.6 | 1.7 | 1.3 | 1.0 | 1.4 | 0.7 | 1.0 |
| %Effectiveness calibration data | 100 | 97 | 100 | 100 | 100 | 100 | 100 | 100 |
| %Effectiveness test data | 97 | 97 | 97 | 97 | 100 | 98 | 100 | 100 |

*PP1 = SNV + SavGol[23,2,2], PP2 = Icoshift + SavGol(25,0,0) + Normalization PP3 = PP1 & PP2 on each data block*
*PP4 = VSN + SavGol[25,1,1], PP5 = PP4 & PP3 on each data block*

*Table 3 Appendix 6. Diesel pour point modelling results*

| Variable predicted | **Pour Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±6 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Regression Method | PLS | | | | PLS + MLR | | | |
| Pre-processing scheme | PP1 | | PP2 | PP3 | PP4 | | PP5 | |
| Data fusion level | - | - | - | Mid | High | High | High | High |
| EPO components | - | 3 | - | - | - | - | - | - |
| Calibration Data Points | 54 | 102 | 36 | 36 | 36 | 36 | 36 | 36 |
| Test Data Points | 29 | 29 | 17 | 17 | 17 | 17 | 17 | 17 |
| Total Data Points | 83 | 131 | 53 | 53 | 53 | 53 | 53 | 53 |
| NIR Variables | 2180 | 2180 | - | 6 LVs scores | ŷ 6LVs PLS | ŷ 5LVs PLS | ŷ 3LVs PLS | ŷ 7LVs PLS |
| NMR Variables | - | - | 13926 | 6 LVs scores | - | - | ŷ 7LVs PLS | ŷ 5LVs PLS |
| Process Variables | - | - | - | - | ŷ MLR | ŷ MLR | ŷ MLR | ŷ MLR |
| Latent Variables | 3 | 5 | 4 | 8 | - | - | - | - |
| Calibration $R^2$ | 0.840 | 0.907 | 0.733 | 0.951 | 0.951 | 0.925 | 0.965 | 0.962 |
| Cross Validation $R^2$ | 0.598 | 0.813 | 0.624 | 0.900 | 0.942 | 0.917 | 0.959 | 0.955 |
| Prediction $R^2$ | 0.692 | 0.741 | 0.826 | 0.879 | 0.888 | 0.863 | 0.936 | 0.898 |
| RMSEC | 4.2 | 3.9 | 5.3 | 2.3 | 2.3 | 2.8 | 1.9 | 2.0 |
| RMSECV | 6.6 | 5.4 | 6.5 | 3.3 | 2.5 | 3.0 | 2.1 | 2.2 |
| RMSEP | 5.6 | 5.3 | 4.3 | 3.8 | 3.4 | 3.8 | 2.6 | 3.4 |
| %Effectiveness calibration data | 80 | 88 | 78 | 97 | 100 | 97 | 100 | 100 |
| %Effectiveness test data | 76 | 76 | 88 | 88 | 94 | 94 | 100 | 94 |

*PP1 = VSN + SavGol[13,4,4], PP2 = Icoshift + SavGol(25,0,0) + Normalization, PP3 = PP1 & PP2 on each data block,*
*PP4 = SNV + SavGol[11,3,3], PP5 = PP4 & PP3 on each data block*

*Table 4 Appendix 6. Diesel cloud point modelling results*

| Variable predicted | **Cloud Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±4 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Regression Method | PLS | | | | PLS + MLR | | | |
| Pre-processing scheme | PP1 | | PP2 | PP3 | PP4 | | PP5 | |
| Data fusion level | - | - | - | High | High | High | High | High |
| EPO components | - | 10 | - | - | - | - | - | - |
| Calibration Data Points | 76 | 140 | 51 | 51 | 51 | 51 | 51 | 51 |
| Test Data Points | 30 | 30 | 23 | 23 | 23 | 23 | 23 | 23 |
| Total Data Points | 106 | 170 | 74 | 74 | 74 | 74 | 74 | 74 |
| NIR Variables | 2180 | 2180 | - | ŷ 11LVs PLS | ŷ 10LVs PLS | ŷ 9LVs PLS | ŷ 3LVs PLS | ŷ 7LVs PLS |
| NMR Variables | - | - | 13926 | ŷ 9LVs PLS | - | - | ŷ 7LVs PLS | ŷ 5LVs PLS |
| Process Variables | - | - | - | - | ŷ MLR | ŷ MLR | ŷ MLR | ŷ MLR |
| Latent Variables | 4 | 8 | 3 | - | - | - | - | - |
| Calibration $R^2$ | 0.760 | 0.936 | 0.460 | 0.914 | 0.850 | 0.809 | 0.922 | 0.915 |
| Cross Validation $R^2$ | 0.678 | 0.808 | 0.413 | 0.908 | 0.842 | 0.802 | 0.880 | 0.906 |
| Prediction $R^2$ | 0.758 | 0.727 | 0.723 | 0.740 | 0.495 | 0.388 | 0.829 | 0.752 |
| RMSEC | 3.8 | 2.2 | 4.9 | 2.0 | 2.6 | 2.9 | 1.9 | 2.0 |
| RMSECV | 4.4 | 3.8 | 5.2 | 2.0 | 2.7 | 3.0 | 2.0 | 2.1 |
| RMSEP | 3.0 | 3.3 | 3.3 | 3.3 | 4.0 | 4.4 | 2.8 | 2.9 |
| %Effectiveness calibration data | 72 | 94 | 73 | 96 | 90 | 84 | 94 | 96 |
| %Effectiveness test data | 80 | 77 | 74 | 91 | 83 | 78 | 83 | 87 |

*PP1 = VSN + SavGol[13,4,3], PP2 = Icoshift + SavGol(25,0,0) + Normalization, PP3 = PP1 & PP2 on each data block,*
*PP4 = SNV + SavGol[25,2,2], PP5 = PP4 & PP3 on each data block*

*Table 5 Appendix 6. Diesel cold filter plugging point modelling results*

| Variable predicted | **Cold Filter Plugging Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±3-0.06*CFPP | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Regression Method | PLS | | | | PLS + MLR | | | |
| Pre-processing scheme | PP1 | | PP2 | PP3 | PP4 | | PP5 | |
| Data fusion level | - | - | - | Mid | High | High | High | High |
| EPO components | - | 2 | - | - | - | - | - | - |
| Calibration Data Points | 65 | 121 | 40 | 40 | 40 | 40 | 40 | 40 |
| Test Data Points | 30 | 30 | 18 | 18 | 18 | 18 | 18 | 18 |
| Total Data Points | 95 | 151 | 58 | 58 | 58 | 58 | 58 | 58 |
| NIR Variables | 2180 | 2180 | - | 6 LVs scores | ŷ 6LVs PLS | ŷ 6LVs PLS | ŷ 6LVs PLS | ŷ 7LVs PLS |
| NMR Variables | - | - | 13926 | 6 LVs scores | - | - | ŷ 6LVs PLS | ŷ 5LVs PLS |
| Process Variables | - | - | - | - | ŷ MLR | ŷ MLR | ŷ MLR | ŷ MLR |
| Latent Variables | 9 | 5 | 4 | - | - | - | - | - |
| Calibration $R^2$ | 0.820 | 0.889 | 0.856 | 0.971 | 0.960 | 0.961 | 0.967 | 0.977 |
| Cross Validation $R^2$ | 0.687 | 0.792 | 0.685 | 0.945 | 0.954 | 0.954 | 0.961 | 0.973 |
| Prediction $R^2$ | 0.714 | 0.742 | 0.849 | 0.925 | 0.918 | 0.909 | 0.931 | 0.936 |
| RMSEC | 3.2 | 3.3 | 3.1 | 1.4 | 1.6 | 1.6 | 1.5 | 1.2 |
| RMSECV | 4.3 | 3.9 | 4.6 | 1.9 | 1.7 | 1.7 | 1.6 | 1.3 |
| RMSEP | 3.9 | 3.8 | 3.1 | 2.2 | 2.2 | 2.33 | 2.1 | 2.1 |
| %Effectiveness calibration data | 78 | 78 | 82 | 100 | 100 | 98 | 100 | 100 |
| %Effectiveness test data | 77 | 47 | 78 | 94 | 94 | 94 | 100 | 94 |

*PP1 = VSN + EMSC, PP2 = Icoshift + SavGol(25,0,0) + Normalization, PP3 = PP1 & PP2 on each data block, PP4 = VSN + SavGol[9,4,3], PP5 = PP4 & PP3 on each data block*

*Table 6 Appendix 6. Kerosene cetane number modelling results*

| Variable predicted | **Cetane Number** | |
|---|---|---|
| Reference method reproducibility | ±3.6 | |
| Model | 1 | 2 |
| Regression Method | PLS | |
| Pre-processing scheme | PP1 | |
| Data fusion level | - | - |
| EPO components | - | |
| Calibration Data Points | 61 | 97 |
| Test Data Points | 29 | 29 |
| Total Data Points | 90 | 126 |
| NIR Variables | 2180 | 2180 |
| NMR Variables | - | - |
| Process Variables | - | - |
| Latent Variables | 9 | - |
| Calibration $R^2$ | 0.986 | 0.991 |
| Cross Validation $R^2$ | 0.97 | 0.968 |
| Prediction $R^2$ | 0.964 | 0.946 |
| RMSEC | 0.6 | 0.5 |
| RMSECV | 1.0 | 1.0 |
| RMSEP | 0.6 | 0.7 |
| %Effectiveness calibration data | 100 | 100 |
| %Effectiveness test data | 100 | 100 |

*PP1 = SNV + SavGol [23,3,2]*

*Table 7 Appendix 5. Kerosene flash point modelling results*

| Variable predicted | **Flash Point** |
|---|---|
| Reference method reproducibility | ±0.071*FP |
| Model | 1 |
| Regression Method | PLS |
| Pre-processing scheme | PP1 |
| Data fusion level | - |
| EPO components | - |
| Calibration Data Points | 41 |
| Test Data Points | 20 |
| Total Data Points | 61 |
| NIR Variables | 2180 |
| NMR Variables | - |
| Process Variables | - |
| Latent Variables | 8 |
| Calibration $R^2$ | 0.836 |
| Cross Validation $R^2$ | 0.613 |
| Prediction $R^2$ | 0.659 |
| RMSEC | 1.3 |
| RMSECV | 2.1 |
| RMSEP | 1.9 |
| %Effectiveness calibration data | 100 |
| %Effectiveness test data | 100 |

*PP1 = SNV + SavGol [13,3,3]*

*Table 8 Appendix 6. Kerosene smoke point modelling results*

| Variable predicted | **Smoke Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±3 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Regression Method | PLS | | | | PLS + MLR | | | |
| Pre-processing scheme | PP1 | | PP2 | PP3 | PP4 | | PP5 | |
| Data fusion level | - | - | - | Mid | High | High | High | High |
| EPO components | - | 3 | - | - | - | - | - | - |
| Calibration Data Points | 44 | 84 | 44 | 44 | 44 | 44 | 44 | 44 |
| Test Data Points | 23 | 23 | 23 | 23 | 23 | 23 | 23 | 23 |
| Total Data Points | 67 | 107 | 67 | 67 | 67 | 67 | 67 | 67 |
| NIR Variables | 2180 | 2180 | - | 6 LVs scores | ŷ 6LVs PLS | ŷ 5LVs PLS | ŷ 5LVs PLS | ŷ 5LVs PLS |
| NMR Variables | - | - | 13926 | 8 LVs scores | - | - | ŷ 5LVs PLS | ŷ 5LVs PLS |
| Process Variables | - | - | - | - | ŷ MLR | ŷ MLR | ŷ MLR | ŷ MLR |
| Latent Variables | 5 | 6 | 3 | - | - | - | - | - |
| Calibration $R^2$ | 0.903 | 0.923 | 0.780 | 0.902 | 0.912 | 0.903 | 0.915 | 0.914 |
| Cross Validation $R^2$ | 0.769 | 0.871 | 0.640 | 0.877 | 0.909 | 0.895 | 0.909 | 0.905 |
| Prediction $R^2$ | 0.722 | 0.735 | 0.589 | 0.734 | 0.740 | 0.723 | 0.732 | 0.721 |
| RMSEC | 1.2 | 1.3 | 1.8 | 1.2 | 1.1 | 1.2 | 1.1 | 1.1 |
| RMSECV | 1.8 | 1.7 | 2.3 | 1.4 | 1.2 | 1.2 | 1.2 | 1.2 |
| RMSEP | 1.9 | 2.1 | 2.5 | 1.9 | 1.8 | 1.9 | 1.9 | 1.9 |
| %Effectiveness calibration data | 96 | 96 | 96 | 96 | 96 | 96 | 96 | 96 |
| %Effectiveness test data | 87 | 87 | 83 | 91 | 91 | 87 | 96 | 87 |

*PP1 = VSN + SavGol[9,4,3], PP2 = Icoshift + SavGol(25,0,0) + Normalization, PP3 = PP1 & PP2 on each data block,*
*PP4 = VSN + SavGol[9,4,3], PP5 = PP4 & PP3 on each data block*

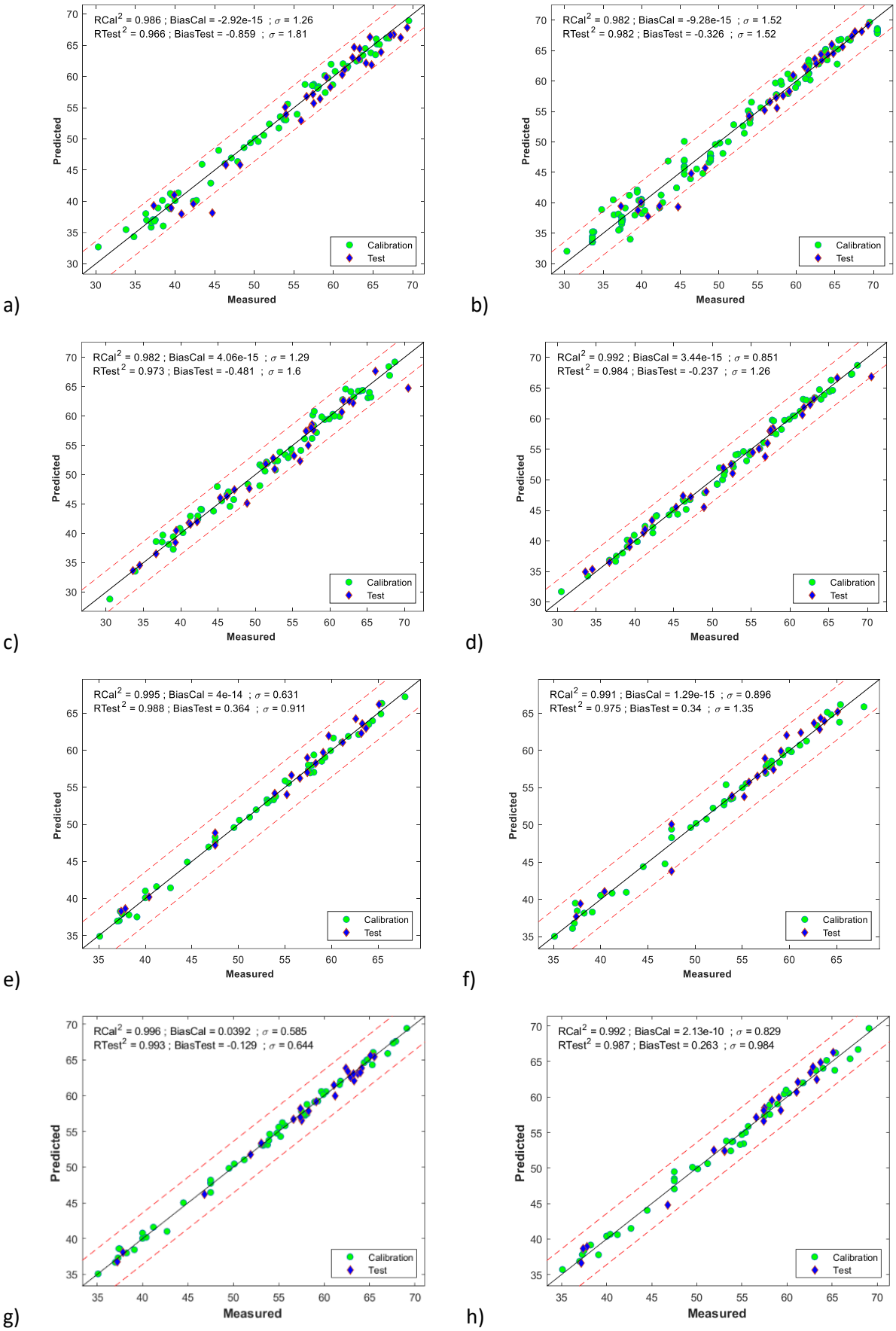



*Figure 1 Appendix 6. Diesel cetane number model parity charts. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

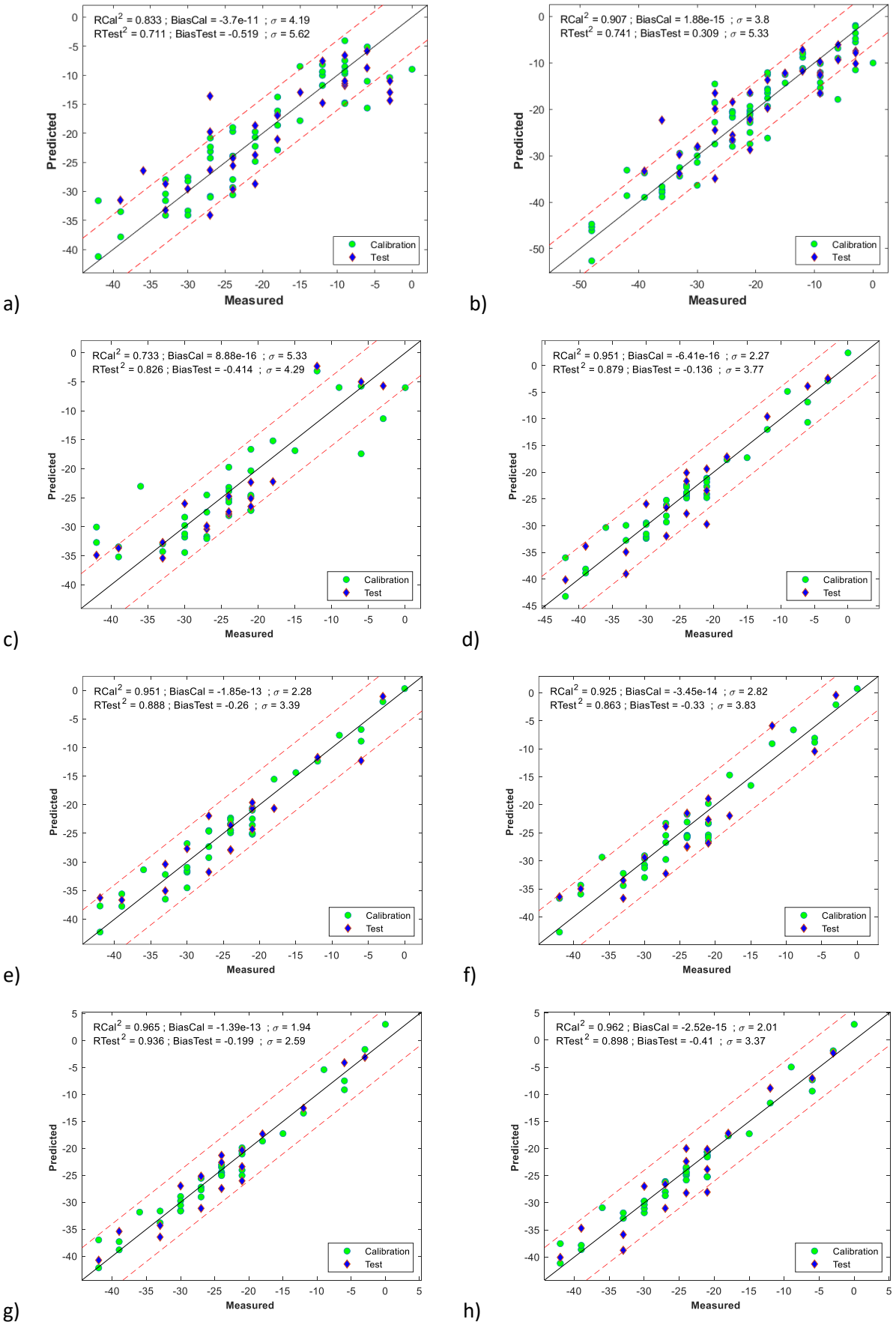



*Figure 2 Appendix 6. Diesel pour point model parity charts. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

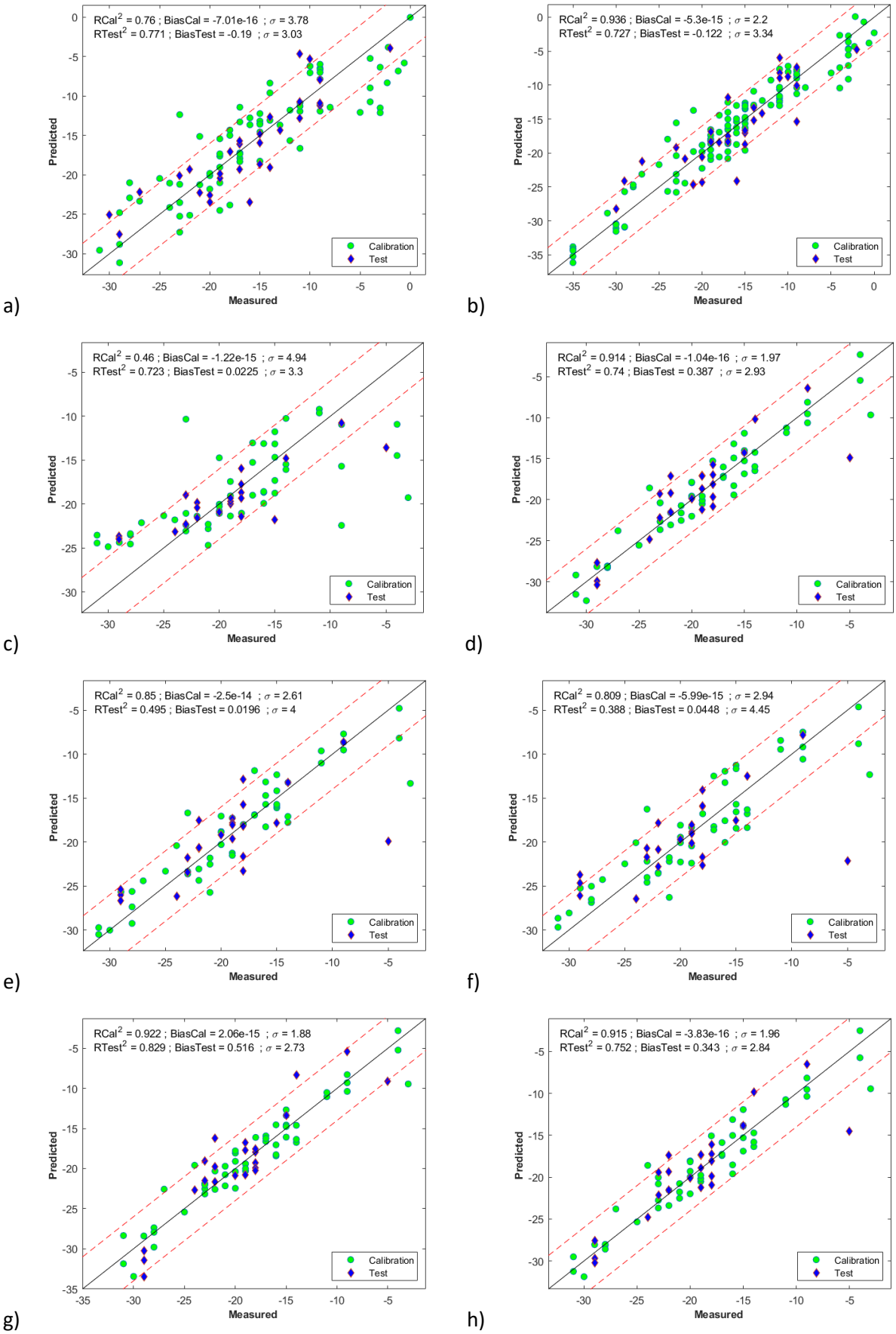



*Figure 3 Appendix 6. Diesel cloud point model parity charts. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

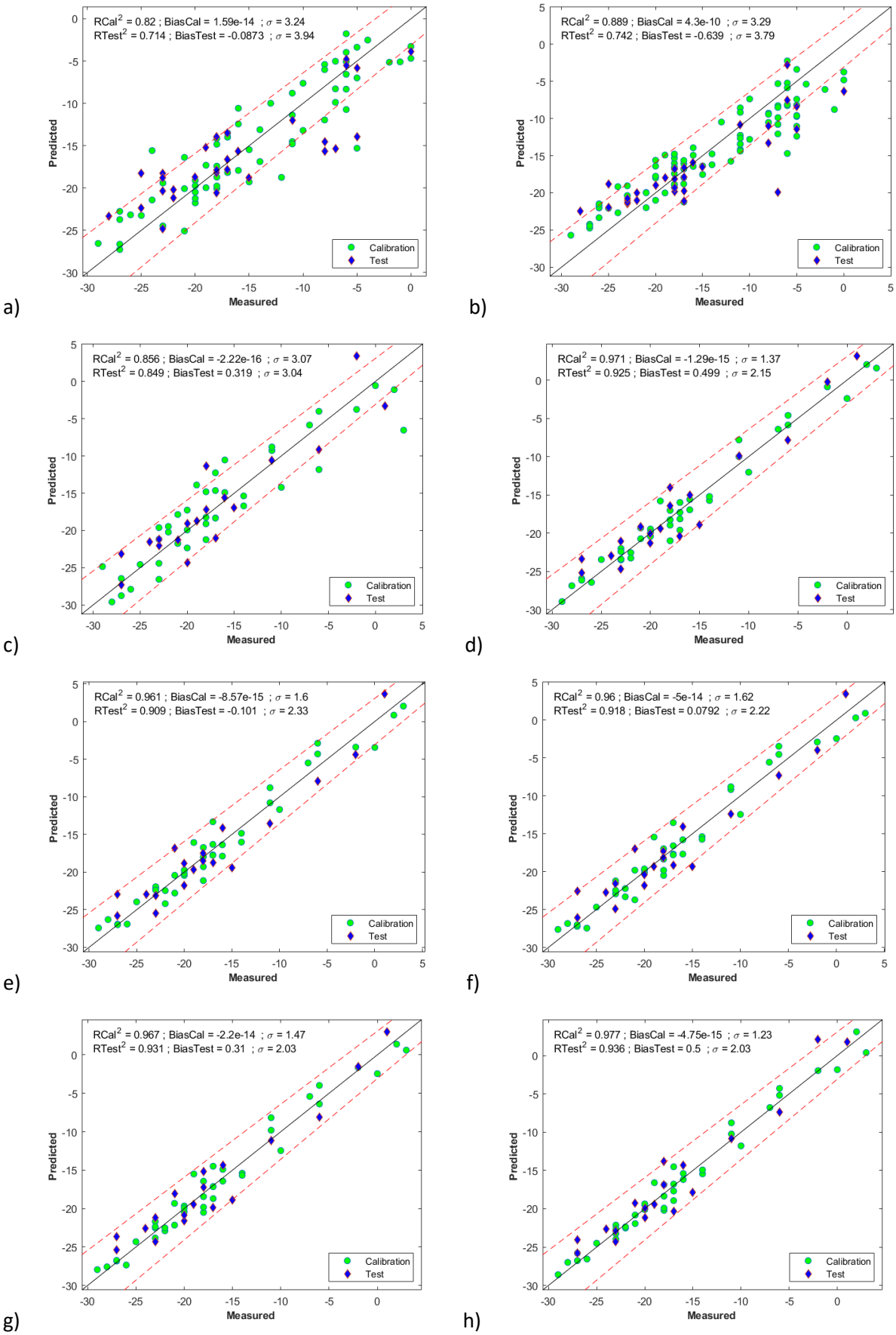



*Figure 4 Appendix 6. Diesel cold filter plugging point model parity charts. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

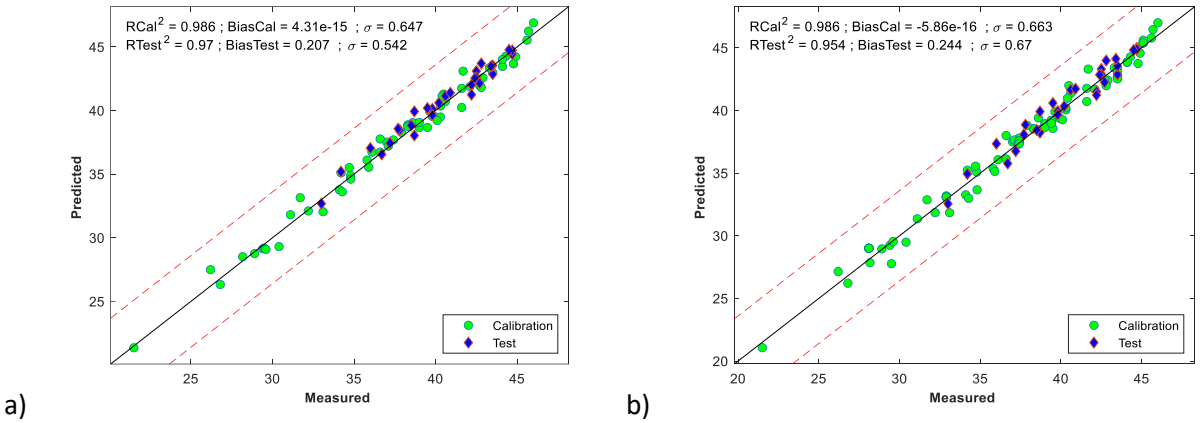



*Figure 5 Appendix 6. Kerosene cetane number model parity charts. a) Model 1, b) Model 2*

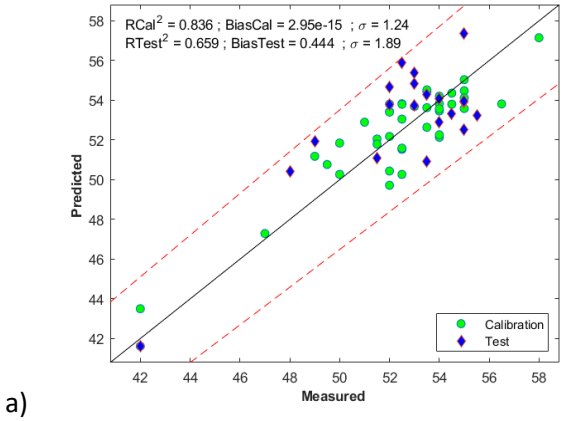



*Figure 6 Appendix 6. Kerosene flash point model parity charts. a) Model 1*

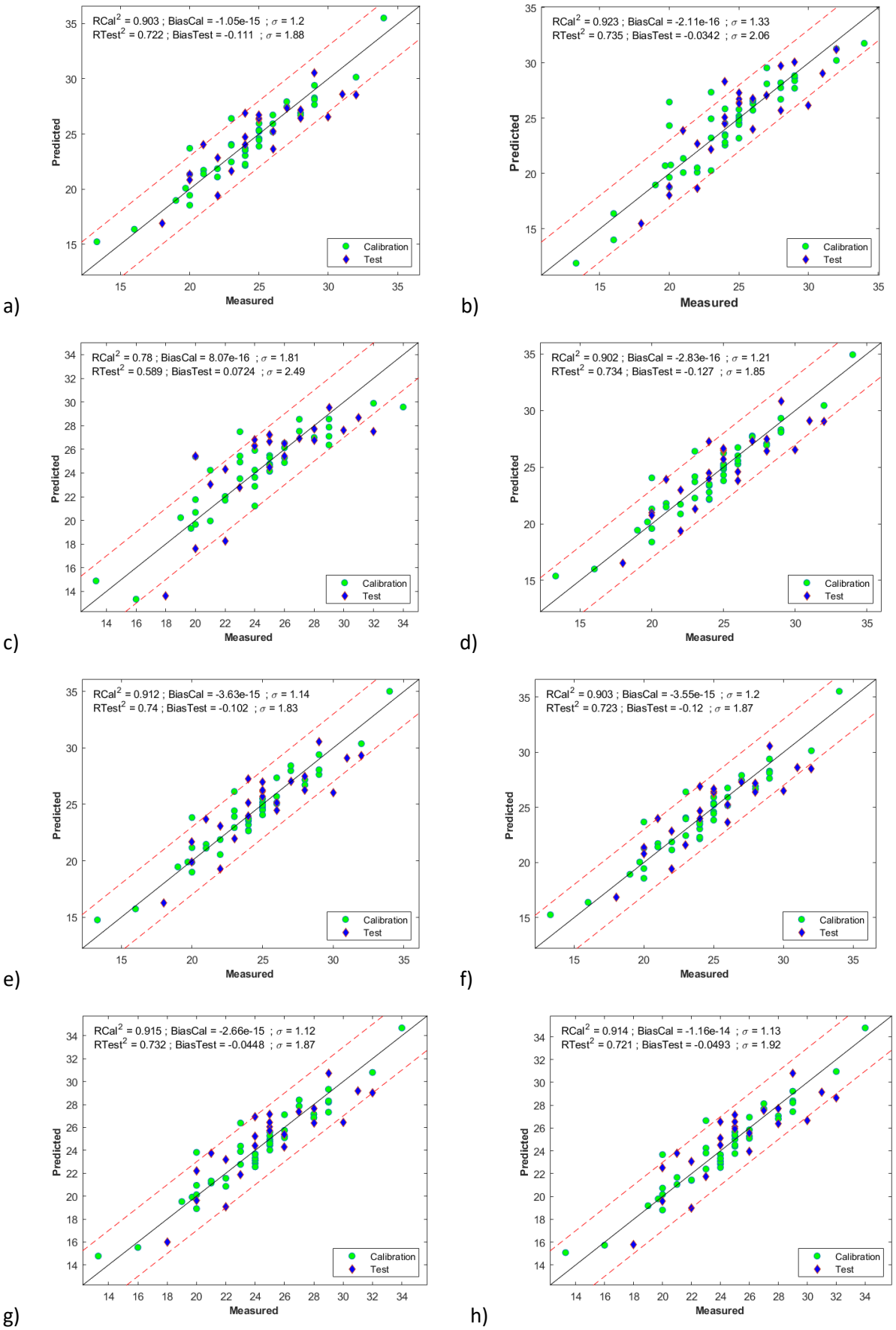



*Figure 7 Appendix 6. Kerosene smoke point model parity charts. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

# Appendix 7: Detailed models validation results

*Table 1 Appendix 7. Description of the models*

| Model | Description | Acronym |
|---|---|---|
| 1 | NIR model | Mod_1 |
| 2 | NIR model with EPO correcttion | Mod_2 |
| 3 | NMR model | Mod_3 |
| 4 | NIR + NMR data fusion model | Mod_4 |
| 5 | NIR + PV_TE* data fusion model | Mod_5 |
| 6 | NIR + PV_NTE** data fusion model | Mod_6 |
| 7 | NIR + NMR + PV_TE data fusion model | Mod_7 |
| 8 | NIR + NMR + PV_NTE data fusion model | Mod_8 |

**PV_TE = Process Variables including Total Effluente properties*
**PV_NTE = Process Variables not including Total Effluente properties*

*Table 2 Appendix 7. Diesel cetane number validation results with 26 new samples*

| Variable predicted | **Cetane Number** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±3.6 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Samples used in test | 26 | | | | | | | |
| Prediction $R^2$ | 0.928 | 0.946 | 0.904 | 0.901 | 0.972 | 0.964 | 0.955 | 0.948 |
| Prediction Bias | 0.3 | -0.4 | 1.4 | 0.0 | 0.2 | 0.6 | -0.3 | 0.6 |
| SEP | 2.2 | 1.1 | 1.6 | 2.0 | 0.7 | 0.9 | 0.9 | 1.0 |
| RMSEP | 2.2 | 1.1 | 2.1 | 2.0 | 0.8 | 1.1 | 1.0 | 1.2 |
| Samples predicted within reproducibility limits | 24 | 26 | 26 | 25 | 26 | 26 | 26 | 26 |
| %Effectiveness test data | 92 | 100 | 100 | 96 | 100 | 100 | 100 | 100 |

*Table 3 Appendix 7. Diesel pour point validation results with 26 new samples*

| Variable predicted | **Pour Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±6 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Samples used in test | 26 | | | | | | | |
| Prediction $R^2$ | 0.455 | 0.503 | 0.572 | 0.629 | 0.542 | 0.476 | 0.525 | 0.496 |
| Prediction Bias | 2.8 | 2.1 | 0.7 | -2.1 | -0.6 | -1.4 | 0.2 | 0.7 |
| SEP | 4.7 | 2.7 | 5.4 | 2.7 | 2.8 | 2.5 | 2.1 | 2.2 |
| RMSEP | 5.5 | 3.4 | 5.4 | 3.4 | 2.8 | 2.9 | 2.1 | 2.3 |
| Samples predicted within reproducibility limits | 17 | 24 | 16 | 25 | 24 | 24 | 26 | 26 |
| %Effectiveness test data | 65 | 92 | 62 | 96 | 92 | 92 | 100 | 100 |

*Table 4 Appendix 7. Diesel cloud point validation results with 26 new samples*

| Variable predicted | **Cloud Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±4 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Samples used in test | 26 | | | | | | | |
| Prediction $R^2$ | 0.033 | 0.206 | 0.221 | 0.256 | 0.595 | 0.607 | 0.619 | 0.615 |
| Prediction Bias | 1.8 | 1.2 | -0.9 | -0.1 | 0.0 | -0.1 | -0.4 | -0.4 |
| SEP | 2.8 | 2.5 | 2.2 | 2.6 | 1.7 | 1.7 | 1.6 | 1.6 |
| RMSEP | 3.3 | 2.8 | 2.4 | 2.6 | 1.7 | 1.7 | 1.7 | 1.7 |
| Samples predicted within reproducibility limits | 24 | 22 | 24 | 25 | 26 | 26 | 26 | 26 |
| %Effectiveness test data | 92 | 85 | 92 | 96 | 100 | 100 | 100 | 100 |

*Table 5 Appendix 7. Diesel cold filter plugging point validation results with 26 new samples*

| Variable predicted | **Cold Filter Plugging Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±3-0.06*CFPP | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Samples used in test | 26 | | | | | | | |
| Prediction $R^2$ | 0.279 | 0.466 | 0.378 | 0.541 | 0.816 | 0.764 | 0.883 | 0.850 |
| Prediction Bias | 2.1 | -0.5 | 1.3 | 0.3 | 0.9 | 1.9 | 0.1 | 0.9 |
| SEP | 2.9 | 2.5 | 2.8 | 2.7 | 1.5 | 1.6 | 1.2 | 1.4 |
| RMSEP | 3.6 | 2.6 | 3.1 | 2.7 | 1.7 | 2.5 | 1.2 | 1.7 |
| Samples predicted within reproducibility limits | 19 | 21 | 17 | 21 | 25 | 23 | 25 | 25 |
| %Effectiveness test data | 73 | 81 | 65 | 81 | 96 | 88 | 96 | 96 |

*Table 6 Appendix 7. Kerosene smoke point validation results with 26 new samples*

| Variable predicted | **Smoke Point** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Reference method reproducibility | ±3 | | | | | | | |
| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Samples used in test | 18 | | | | | | | |
| Prediction $R^2$ | 0.785 | 0.516 | 0.653 | 0.800 | 0.783 | 0.757 | 0.831 | 0.830 |
| Prediction Bias | -1.5 | -1.1 | -1.0 | -0.8 | -0.1 | -1.0 | -0.1 | 0.5 |
| SEP | 1.1 | 1.3 | 1.2 | 1.0 | 1.1 | 1.0 | 1.0 | 1.0 |
| RMSEP | 1.9 | 1.7 | 1.5 | 1.3 | 1.1 | 1.4 | 1.0 | 1.1 |
| Samples predicted within reproducibility limits | 16 | 17 | 17 | 18 | 18 | 17 | 18 | 18 |
| %Effectiveness test data | 89 | 94 | 94 | 100 | 100 | 94 | 100 | 100 |

*Table 7 Appendix 7. Kerosene cetane number validation results with 26 new samples*

| Variable predicted | **Cetane Number** | |
|---|---|---|
| Reference method reproducibility | ±3.6 | |
| Model | 1 | 2 |
| Samples used in test | 26 | |
| Prediction $R^2$ | 0.922 | 0.928 |
| Prediction Bias | -0.1 | 0.0 |
| SEP | 0.7 | 0.7 |
| RMSEP | 0.7 | 0.7 |
| Samples predicted within reproducibility limits | 26 | 26 |
| %Effectiveness test data | 100 | 100 |

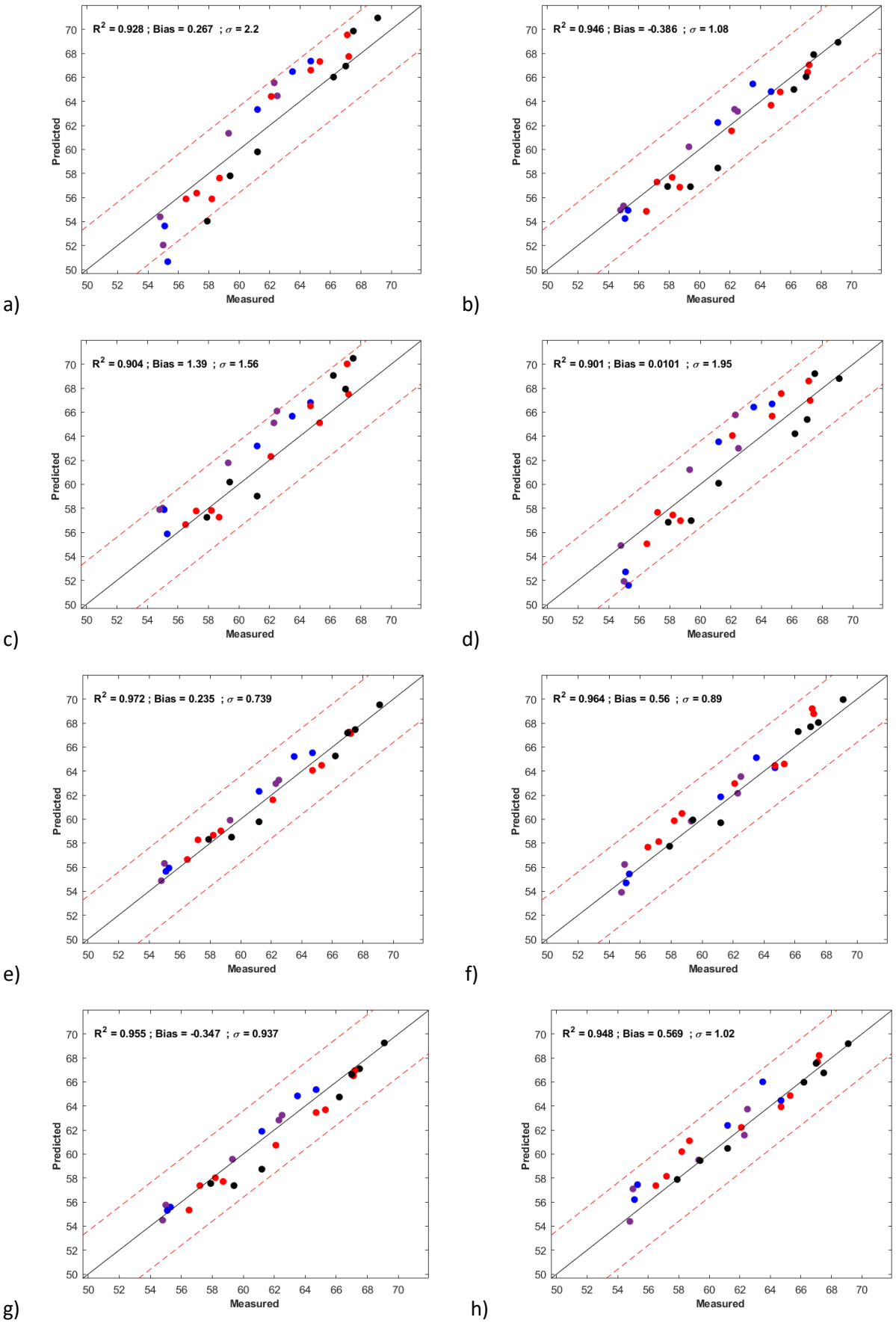



*Figure 1 Appendix 7. Parity charts of diesel cetane number validation with 26 new samples. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

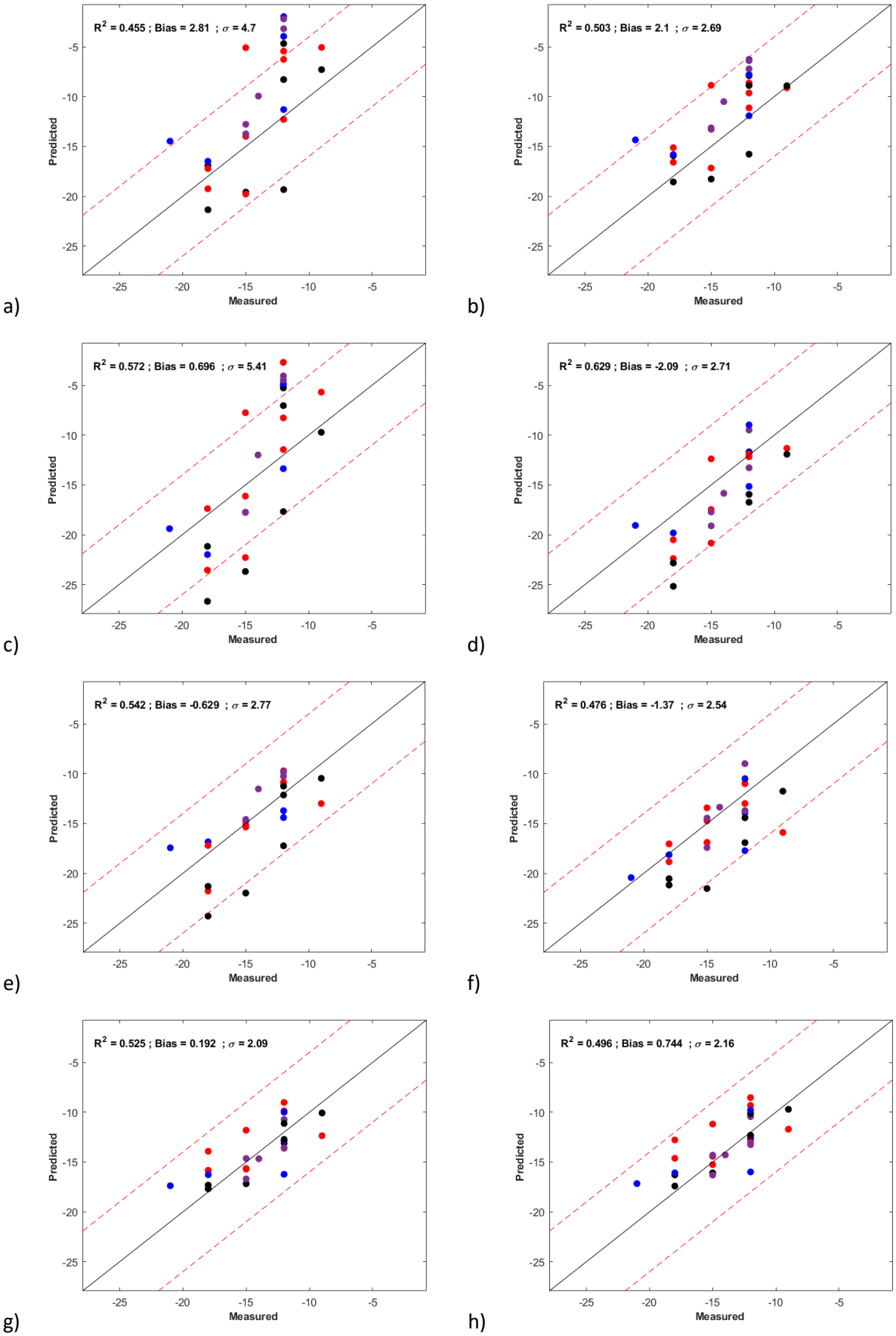



*Figure 2 Appendix 7. Parity charts of diesel pour point validation with 26 new samples. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

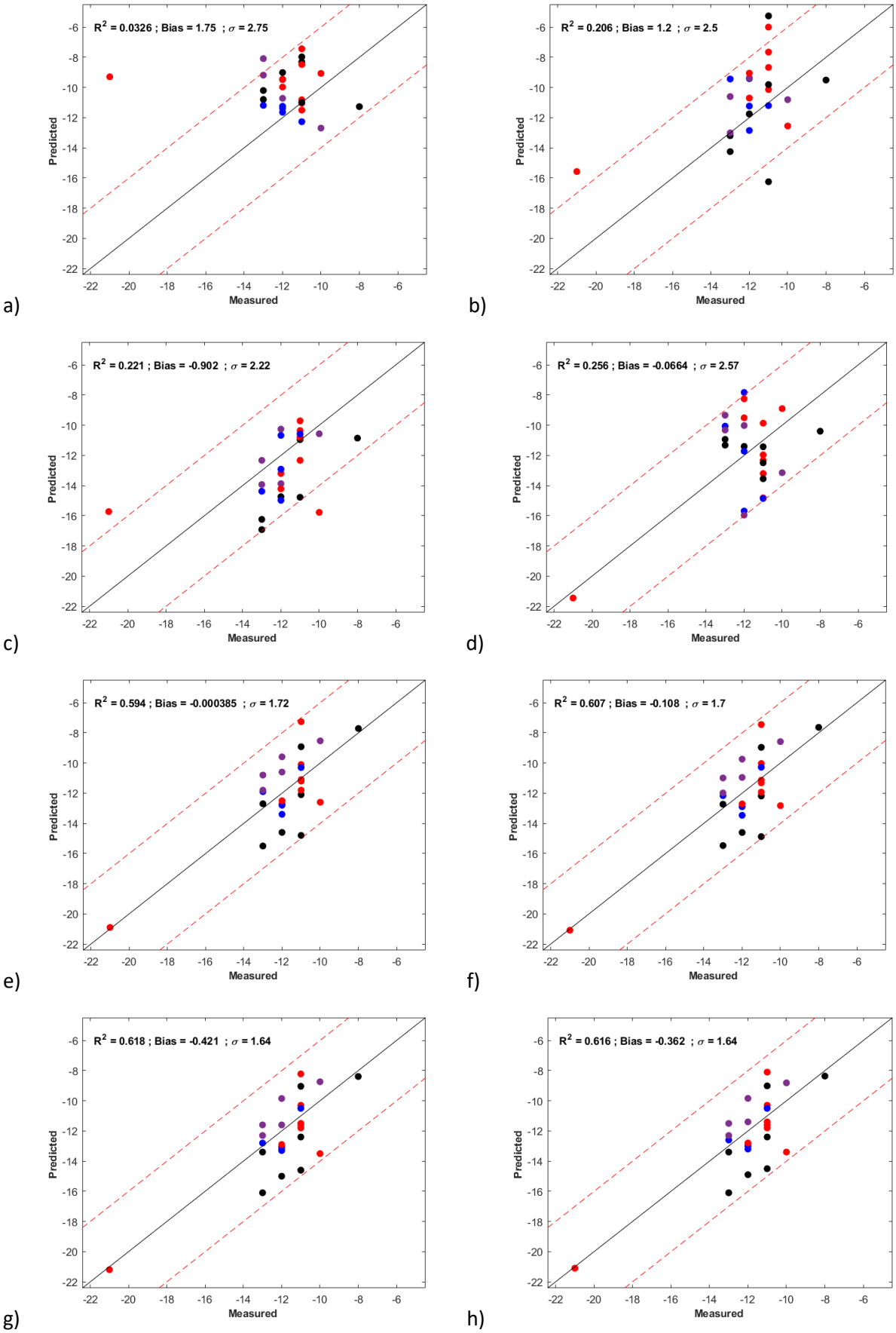



*Figure 3 Appendix 7. Parity charts of diesel cloud point validation with 26 new samples. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

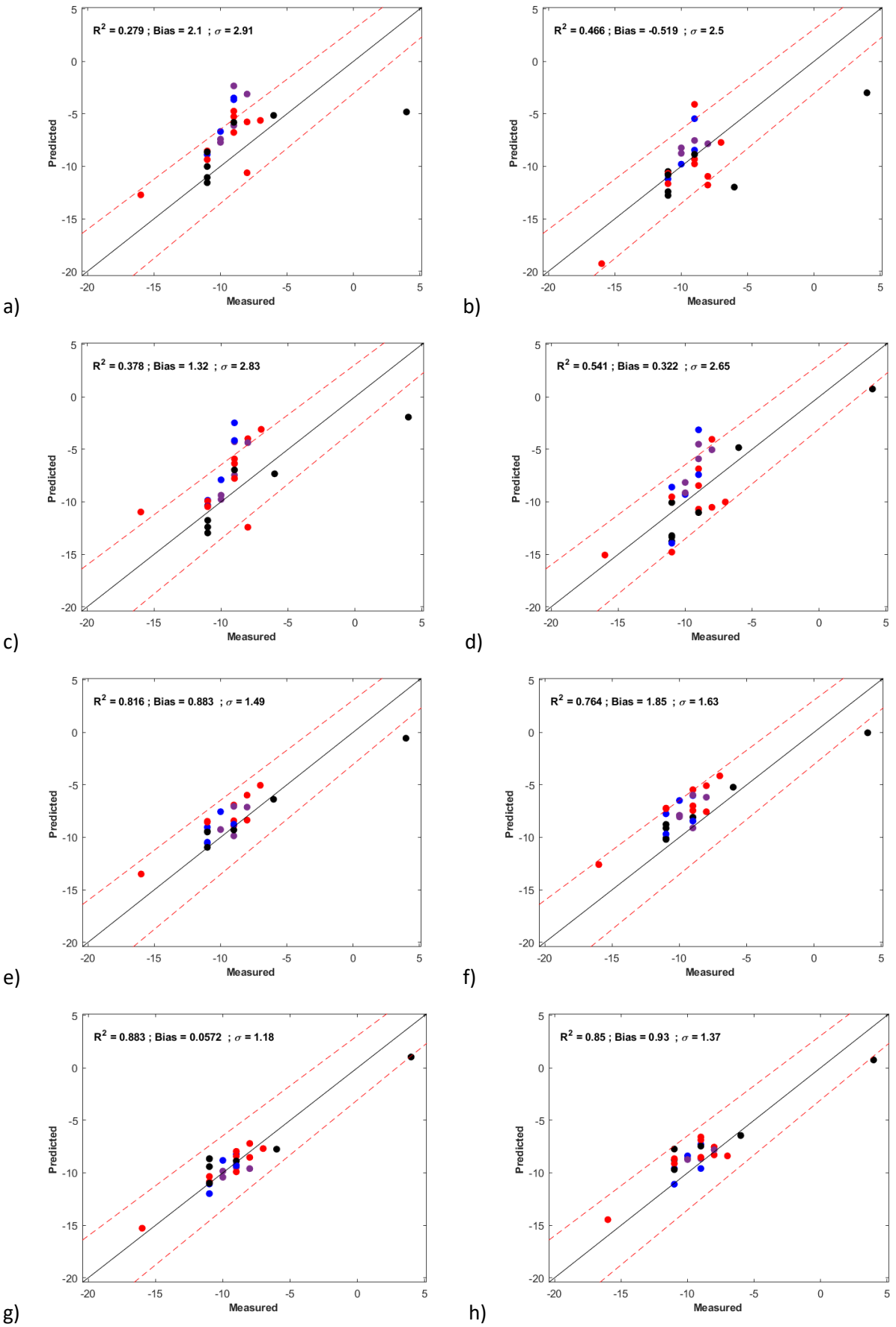



*Figure 4 Appendix 7. Parity charts of diesel cold filter plugging point validation with 26 new samples. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

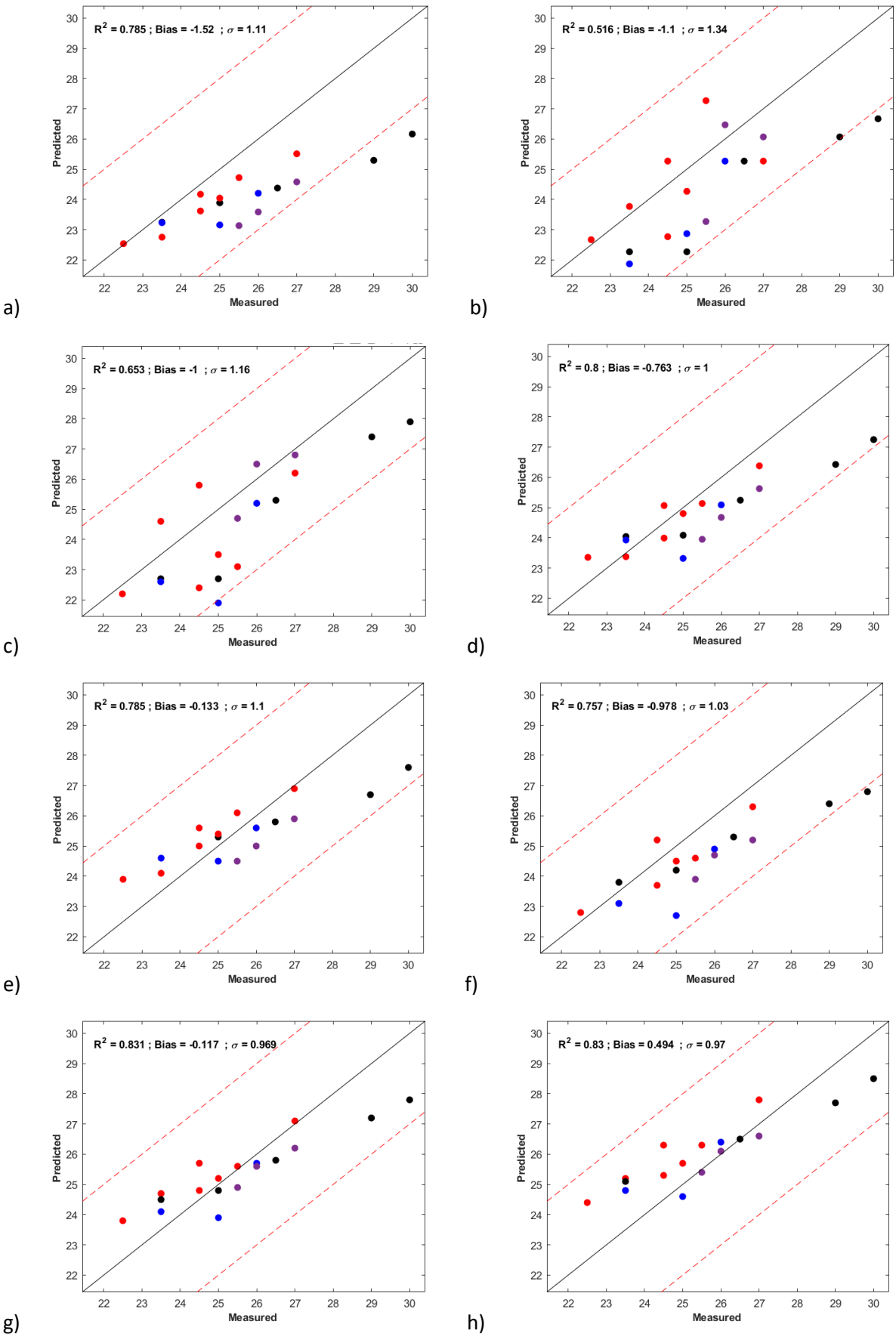
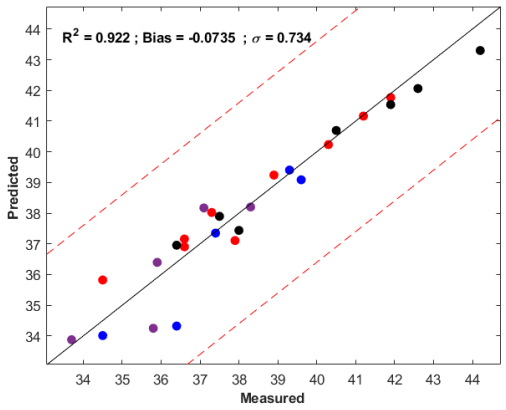



*Figure 5 Appendix 7. Parity charts of kerosene smoke point validation with 26 new samples. a) Model 1, b) Model 2, c) Model 3, d) Model 4, e) Model 5, f) Model 6, g) Model 7, h) Model 8*

a)




b)

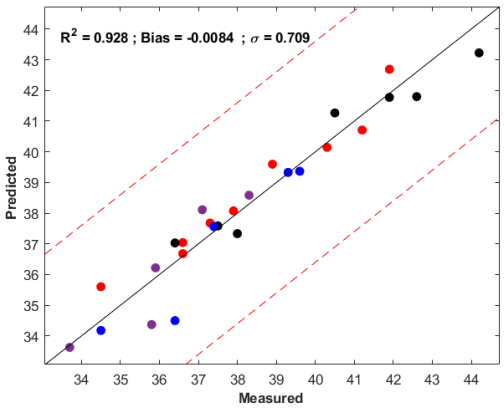



*Figure 6 Appendix 7. Parity charts of kerosene cetane number validation with 26 new samples. a) Model 1, b) Model 2*

# Appendix 8: Supplementary theoretical information

**Preprocessing methods**

There is substantial literature on spectroscopic modelling applied to different industries where preprocessing is an integral part of the studies. In the case of vibrational spectroscopy, there are several preprocessing techniques[164] since spectrum quality can be affected by systemic noise, such as baseline variation and multiplicative effects. Different mathematical methods can be used to remove the noisy information from the spectral signal. However, the preprocessing must be carefully done as it can also lead to the removal of relevant information[165]. There is no single procedure or routine established for data preprocessing. This task is far from being unique and simple, as it depends on the database characteristics, data source, nature of the sample and the modelling objective[166]. The most common approach for selecting the preprocessing method is to evaluate the information retained after preprocessing that best describes the studied property.

Preprocessing methods for spectroscopic data can be divided into two main groups according to the causative effect of the noise: dispersion correction methods (scatter-corrective) and spectral derivatives[164]. The main difference between these two groups is that the scatter correction methods are used to reduce the scatter effects occurring between samples, whereas derivatives methods are employed to remove baseline effects and any offset differences between the data. Table App 8.1 shows the most used preprocessing methods according to their group.

*Table App 8.1. Most common preprocessing methods for spectral information [167, 164]*

| Scatter-Corrective | Spectral derivatives |
|---|---|
| Multiplicative Scatter Correction (MSC)[122] | Automatic Weighted Least Squares Baseline (AWLS-B)[124] |
| Extended MSC (EMSC)[125] | Norris-Williams derivation (NW-D)[126] |
| Detrend[121] | Savitzky-Golay derivative (SavGol)[118] |
| Standard Normal Variate (SNV)[121] | Piecewise Direct Standardization (PDS)[150] |
| Variable Sorting for Normalization (VSN)[120] | Loading Space Standardization (LSS)[168] |
| Probabilistic Quotient Normalization (PQN)[123] | |
| Optical Path Lenght Estimation and Correction (OPLEC)[169] | |

**Regression methods**

According to the number of independent variables used in the model development, the regression can be highly multivariate. The higher the number of independent variables used, the greater the complexity of the regression process. In the case of spectral information, the regression is considered highly multivariate since a single spectrum contains a large number of wavelengths (or chemical shifts in the case of the NMR spectra), each of which is an x-variable. Since a spectrum can contain more than 1000 variables, performing a multivariate regression using that amount of information in a conventional approach could be impossible since the X matrix may have more variables than observations. For this reason, it is necessary to reduce the size of the X matrix used without losing relevant information from the evaluated sample. Multivariate analysis using projection techniques on latent variables is a widely practiced procedure to achieve this purpose, being

the Partial Least Square (PLS) the most common.[91]

There are different methods to obtain appropriate and consistent regression models. Selecting the best method depends directly on the nature of the independent variables and their relationship with the dependent variables, which can be null (no dependence), directly or inversely proportional, linear or non-linear. A summary of the most used methods in multivariate regression is presented below in Table App 8.2 y Table App 8.3.

*Table App 8.2. Most common linear multivariate methods for linear regression*

| Method |
|---|
| Classical Least Squares (CLS)[170] |
| Multiple Linear Regression (MLR)[171] |
| Partial Least Squares (PLS)[127] |
| Interval PLS (iPLS)[147] |
| Principal Component Regression (PCR)[113] |

*Table App 8.3. Most common non-linear multivariate methods for linear regression*

| Method |
|---|
| Support Vactor Machine (SVM)[128] |
| Artificial Neural Networks (ANN)[129] |
| Back-Propagation Neural Networks (BPNN)[172] |
| Probabilistic Neural Network (PNN)[173] |
| Random Forest (FR)[174] |
| Projection Pursuit Regression (PPR)[175] |
| Locally Weighted Regression (LWR)[130] |

**Data fusion modelling**

Lahat *et al.*[176] emphasize the need to find the answer to "how to exploit the diversity of information" found in a database. Solving this concern is relevant since the nature, origin, values, and units of measurement of the data integrated directly affect the suitability and veracity of the regression model. Data fusion, whose definition is "the analysis of several data sets such that different data sets can interact and inform each other" [177] seeks to provide an answer to the need addressed.

The data from different sources used in the same regression process (data fusion modelling) must have a common characteristic for their concatenation and integration. In general, there are three structures of information concatenation. (A) Data blocks of the same order, either 2D sharing a common characteristic or 3D sharing two common characteristics, (B) data blocks of the same order 3D sharing one characteristic in common, and (C) data blocks of different order sharing at least one characteristic in common.

Regardless of the data block structure used to perform the data fusion, this modelling approach is characterized by different strategies known as fusion levels (low-, mid-, and high-level)[131]. The low-level fusion consists of using the information from the blocks directly in the development of the model either by

simple concatenation of the blocks or using decomposition or factorization methods on one block regarding another[133]. At the mid-level fusion, a feature extraction step from each dataset is performed first through statistical analyses such as PCA and PLS for their later fusion by simple concatenation[134]. Finally, the high-level fusion combines the decisions or results obtained from developed prediction models separately with each data block [135]. Figure App 8.1 shows a representation of these three levels.

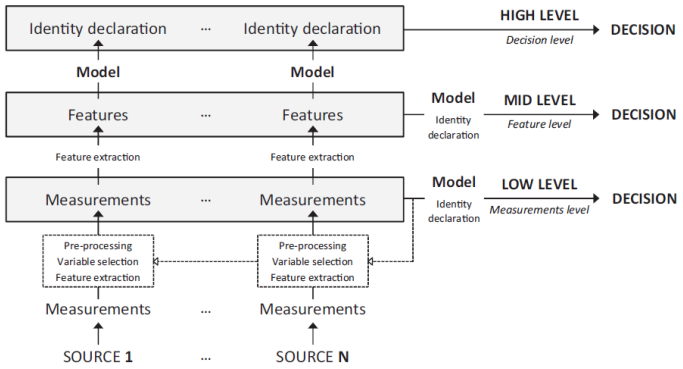



*Figure App 8.1. Data fusion strategies (levels) [178]*

The primary aim of data fusion is to improve regression model performance by using the most relevant characteristics of different information sources. Different authors[179–183] showed in their research that data fusion improves not only model performance, but also the extraction of relevant descriptors of influence. However, the data fusion results depend highly on the nature of the sample analyzed, the fusion techniques employed[184] and the source information used.

**Outlier detection analysis**

Outlier detection methods can be classified into two main categories, unsupervised methods (based on an a priori analysis of available data) and supervised methods (based on an a posteriori analysis following a predefined model). The most common supervised outlier detection methods are leverage effect, Covariance Ratio, and Cook-Distance.

Concerning the unsupervised methods, the most common approach used in the multivariate analysis by projection on latent variables (dimensionality reduction) is the combination of the Q residual test (a measure of the difference between a sample and its projection into the k factors retained in the model) and Hotelling's T-Squared (a measure of the variation in each sample within the model)[185, 186]. An example of using these methods is shown in Figure App 8.2[185], where some data exceeding the threshold of either the Hotelling's T-Squared test (strong outlier) or the Residual Q test (weak outlier) can be observed. This figure also shows a data point exceeding the threshold of both tests. This combined analysis facilitates the identification of potential anomalous data, whether they are of a strong or weak nature. Weak outliers can be excluded from the database if they have significant repercussions on the model. On the other hand, removing strong outliers from the dataset can affect the quality and robustness of the model since their influence on the database is

strong (they may contain relevant descriptors information). If a data point simultaneously exceeds the established thresholds of the two tests, using this information in the model should be reconsidered.

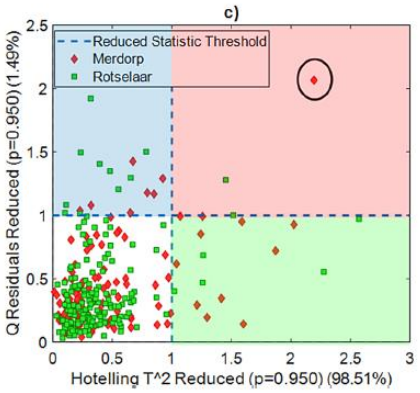



*Figure App 8.2. Data Hotelling T² & Q Residuals Tests*[185]

There is no single solution for detecting and handling outliers [187]. Different methods can be applied depending on the available information and its use in the prediction models developed. The methodology most appropriate is defined by each research study which has to adapt the methods used to its needs.

**Variable selection**

Variable selection in the modelling process can be implemented either for cost reduction in data acquisition (design and sensor selection), optimization in machine consumption, or improvement of model performance, whether related to its accurate prediction or its homoscedasticity in evaluating different parameters affecting the predicted variable. An intrinsic advantage of performing variable selection is a better understanding of the interaction between the independent variables and the estimated variable, which could lead to a better understanding of the evaluated process and its potential optimization.

In 2010 Andersen et al.,[188] conducted a study presenting the use and most common errors committed when applying variable selection methods in highly multivariate data. In this study, four variable selection methods were described: Variable Importance in Projection (VIP)[138], Selectivity Ratio (SR)[139], interval PLS (iPLS)[147], and the Genetic Algorithms (GA)[141].

Some of the most recent developments in the variable selection analysis on highly multivariate information are the works developed by Roger and Biaconlillo et al.[145, 146]. They proposed two methods for variable selection known as Covariance Selection (CovSel) and sequential and orthogonalized CovSel (SO-CovSel).

The variable selection is also applied to low multivariate data. The most methods include the Least Absolute Shrinkage and Selection Operator (LASSO)[140], Recursive Feature Elimination (RFE)[142], and the sequential feature selection (SFS)[143] along with its variant, the sequential floating feature selection method (SFFS)[144].

The application scope of variable selection analysis is quite broad. A comprehensive study related to this subject can be found in the works of Anzanello et al.,[189] and Heinze et al. [190]

**External parameters influence correction**

The spectral variability generated by an external parameter must be corrected, or at least minimized, to ensure a reliable description of the sample physicochemical behavior from the spectroscopic information extracted. Accordingly, Chauchard et al.[149]proposed a general methodology to determine the best strategy to correct the influence caused by an external parameter. In Figure App 8.3, the methodology is summarized, where G is the external parameter and g is its respective value. In summary, when identifying an external parameter with potential impact on the quality of the spectrum, and hence on the model performance, it must be determined whether the influence is significant or negligible. If it is a highly influential parameter, the next step is to determine if it can be controlled. Finally, in case of a negative outcome, it must be established whether the value g of the parameter G is known when using the model. If g is known, there are three strategies for correcting the influence of this parameter. Otherwise, robust modelling must be conducted.

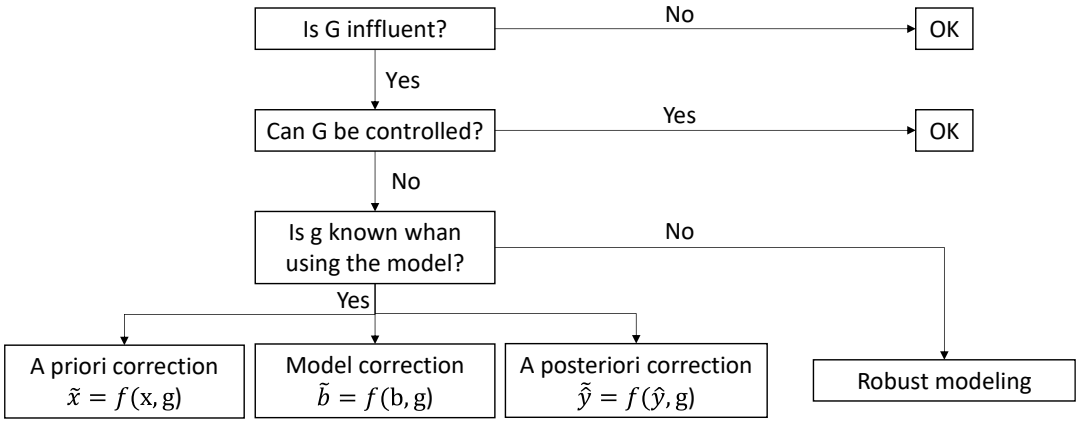



*Figure App 8.3. General methodology for external parameter influence correction[149]*
*Legend: G = external parameter studied, g = value of G, x = x matrix (NIR spectra), b = model slope, y = dependant varuable.*

The a priori correction strategy, as its name indicates, is focused on the correction of the new spectra before it is used. This correction is based on the difference between the existing and the new spectra affected by the G parameter. The most employed methods are the piece direct standardization (PDS)[150] and the spectral space transformation (SST)[191]. Concerning the model correction strategy, local modelling can be taken as an example. In this case, different regression models are generated with consolidated databases at different acquisition conditions defined by g. When a new sample is evaluated, its value is estimated from the sum of the $\hat{y}$ obtained in each developed model, adjusted by a differential factor between the g used in the model development and the g of the new sample. Finally, the most frequent application of the a posteriori correction strategy is the adjustment of the value predicted by correcting the bias and slope, which are affected by the influence of the G parameter.

Compared to the strategies described previously, robust modelling is an alternative having higher efficiency in solving the problem of the constant and sometimes unexpected variability that may occur in the spectra acquisition. In this strategy can be found mainly the orthogonalization methods, such as the external

parameter orthogonalization (EPO)[152] and the dynamic orthogonal projection (DOP)[153]. Another widely used method in this strategy is data augmentation[192].

The definition of the external parameters that can affect the model performance and their respective correction is a task that must be done comprehensively to ensure that the models remain valid over time, and that they can evaluate the influence of the parameters on the estimated properties. This activity becomes even more important when prediction models are intended to be used in real-time process monitoring applications.