Compréhension moléculaire et prédiction des propriétés physicochimiques dans les produits pétroliers

Jean-Jérôme da Costa Soares

N° d'ordre NNT : xxx

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
IFP Energies Nouvelles

Ecole Doctorale ED206
(Ecole doctorale de Chimie de Lyon)

Soutenue publiquement le 14 Décembre 2017, par
Jean-Jérôme Da Costa Soares

Compréhension moléculaire et prédiction des propriétés physico-chimiques dans les produits pétroliers

---

Devant le jury composé de :

RUCKEBUSCH, Cyril	Professeur des Universités / Université de Lille 1, Sciences et Technologies / UMR CNRS 8516 Laboratoire de Spectrochimie Infrarouge et Raman Président
THYBAUT, Joris	Professeur / Université de Gent / Faculté d'ingénierie et d'architecture / Department of Materials, Textiles and Chemical Engineering Laboratory for Chemical Technology Rapporteur
DUPUY, Nathalie	Professeure des Universités / Université Aix-Marseille / IMBE UMR 7263 équipe biotechnologie et chimiométrie Rapporteure
GOURVENEC, Sébastien	Responsable d'équipe Spectroscopie et Modélisation / Département Contrôle Avancé et Analyseurs / TOTAL Raffinage-Chimie Examinateur

BORDES, Claire	Maître de Conférences / Institut des Sciences Analytiques / UMR 5280 / Université Lyon 1 Examinatrice
ESPINAT, Didier	Directeur expert / IFP Energies Nouvelles / Direction Physique et Analyse Directeur de thèse
CELSE, Benoit	Ingénieur de recherche / IFP Energies Nouvelles / Direction Conception Modélisation Procédés Examinateur

Résumé

La diminution en pétrole brut léger nécessite de convertir les fractions lourdes en produits valorisables (essences, gazoles, huiles, etc.). Dans ce contexte, l'hydrocraquage (HCK) fournit des produits de très haute qualité à partir de distillats sous vide (DSV) du pétrole brut. La qualité des coupes obtenues est caractérisée par des propriétés physico-chimiques qui sont soumises à des spécifications. L'optimisation du procédé nécessite des expérimentations longues et coûteuses. IFPEN a donc de plus en plus recours à des tests sur unité d'expérimentation haut débit (EHD). Ces derniers posent cependant un problème d'accessibilité aux coupes d'intérêt. Par ailleurs, pour comprendre et prédire l'impact des conditions opératoires sur la qualité des produits, des simulateurs sont développés. Certaines propriétés de produits sont cependant complexes et difficiles à modéliser voire mal comprises.

Ce travail de thèse a porté sur l'amélioration de la compréhension moléculaire des propriétés produits pour une meilleure prédiction. Dans cette étude, nous nous sommes focalisés sur le point de trouble (PT) de la coupe gazole et l'indice de viscosité (VI) de l'huile obtenue lors de l'hydrocraquage de DSV. Deux techniques d'analyse moléculaire ont été utilisées : la chromatographie en phase gazeuse bidimensionnelle (GC×GC) qui permet de déterminer la composition par famille chimique des différentes coupes et la résonance magnétique nucléaire (RMN) du ^{13}C qui fournit des informations sur la structure chimique des hydrocarbures présents dans ces mélanges. Nous présentons les résultats obtenus par une régression multivariée parcimonieuse (*sparse Partial Least Squares*) appliquée aux données GC×GC et ^{13}C RMN. Il s'agit d'une variante de la PLS classique qui permet de réduire le nombre de facteurs tout en privilégiant ceux qui sont les plus corrélés à une propriété d'intérêt donnée. Globalement, cette étude a notamment permis de mieux comprendre l'impact des différents hydrocarbures (n-paraffines, isoparaffines, aromatiques,…) et de leur structure moléculaire (longueur de chaînes, degrés de branchements,…) sur le PT des gazoles et le VI des huiles. **La bonne qualité des modèles obtenus par *sparse* PLS montre par ailleurs la possibilité d'accéder à la qualité des produits lors de l'utilisation d'EHD**. Des modèles de prédiction par krigeage ont également été développés. Cette méthode d'interpolation permet de prédire une propriété en un point donné en effectuant une moyenne pondérée des observations au voisinage de ce point. Les modèles de krigeage sont des modèles locaux adaptés aux structures de données complexes. Ce sont des approches probabilistes qui permettent d'estimer les incertitudes de prédiction. **Aussi bien dans le cas du PT de la coupe gazole que dans celui du VI de la coupe huile, les résultats montrent une amélioration des performances.** Cette approche est tout à fait novatrice dans le domaine des produits pétroliers. **Lors de l'utilisation d'unités EHD, elle permet d'accéder au VI des huiles de base plus aisément que *via* des données chromatographiques ou spectroscopiques, qui sont de plus non accessibles en raffinerie.**

Abstract

The rapid decline in light crude oils requires to convert heavy petroleum fractions into more valuable products (naphtha, diesel, lubricants, *etc.*). In this context, hydrocracking process (HCK) consists on upgrading vaccum gas oil (VGO) into high quality products. The quality of petroleum products is based on some chemical and physical properties that should fulfill prerequisite specifications. The hydrocracking process optimization requires to set up time consuming and costly experiments for developing catalysts and setting operating conditions. High throughput experimentation (HTE) units are then increasingly used at IFPEN. However, these units do not enable to obtain end products. Otherwise, predictive models were developed in order to understand and predict the impact of operating conditions about products quality. However, some complex properties are very difficult to model and require a better understanding.

This work is mainly concerned with the understanding of diesel cloud point (CP) and viscosity index (VI) of base oils. Two analytical techniques were used: the two-dimensional gas chromatography (GC×GC) that enables to identify hydrocarbons compounds in petroleum products and the ^{13}C nuclear magnetic resonance (NMR) spectroscopy which provides structural characteristics of these compounds. A sparse multivariate regression (sparse Partial Least Squares) was performed using chromatographic and spectroscopic data. The sparse PLS is derived from classical PLS. It allows to reduce the number of factors by performing a variable selection. The selected factors are the most correlated to the property to model. Globally, this approach enabled to better understand how hydrocarbon compounds (n-paraffins, isoparaffins, aromatics,…) and their molecular characteristics (carbon number, degree of branching,…) affect the diesel CP and the VI of base oil. **Furthermore, the good performances of developed sparse PLS models show that it is possible to access to the products quality when using HTE units.** Kriging models were also developed. Kriging is an interpolation method that predicts the value of a function at a given point by computing a weighted average of the known values of the function in the neighborhood of the point. Kriging models have local aspect which is well adapted to complex data. Its probabilistic approach enables to provide an estimate of predicted value uncertainty. **Results show that kriging improves predictive performances for both diesel CP and VI of base oil.** This approach is quite innovative in modelling of petroleum products properties. **When using HTE units, it allows to estimate the VI of base oil more easily than from chromatographic or spectroscopic data which are not available for the refiners.**

Remerciements

Ce travail de thèse a été réalisé au sein de la Direction Physique et Analyse d'IFP Energies Nouvelles de Solaize.

Je tiens à remercier tout particulièrement mon Directeur de Thèse Didier ESPINAT pour ses conseils avisés, son point de vue toujours pertinent et sa bonne humeur. Je te souhaite de bien profiter de ta retraite.

J'adresse mes sincères remerciements à mes deux promoteurs Fabien CHAINET et Benoit CELSE qui m'ont fait confiance durant ces trois années et qui m'ont apporté leur soutien moral et intellectuel à tout moment. Fabien, je te remercie pour ton implication totale dans ces travaux. Ta disponibilité et ta rigueur ont été essentiel pour ma réussite. J'ai une pensée toute particulière pour toi Benoit que je considère comme un véritable mentor et un modèle à suivre.

Je remercie également Marion LACOUE-NEGRE. Bien que tu ne sois pas officiellement un promoteur de cette Thèse, ton implication a été primordiale. Tu as été présente du début à la fin. Plus encore tes travaux sur la résonance magnétique nucléaire m'ont ouvert la voie à l'application de cette technique pour la résolution de problématiques liées à ce sujet de Thèse. Remerciements chaleureux à Cyril RUCKEBUSCH pour son expertise tant sur le fond que sur la forme dans la mise en valeur de ces travaux de Thèse. Ta vision souvent aux antipodes de la mienne m'a permis de me confronter aux exigences du monde de la recherche et donc d'évoluer notamment au niveau de la rigueur. Je n'oublie pas de remercier Noémie CAILLOL qui a été elle aussi présente de bout en bout et avec qui j'ai eu des échanges toujours fructueux.

Je remercie tous ceux qui ont contribué à la réussite de ce projet : Frédéric NEYRET-MARTINEZ, Frédéric FILALI et Lucie BUSSOD à qui je dois notamment la réalisation des analyses en Résonance Magnétique Nucléaire ; Yohann MOUILLET et Marjorie BOIRON qui ont effectué l'ensemble des analyses de chromatographie en phase gazeuse bidimensionnelle ; Jérémy PONTHUS qui m'a apporté ses conseils avisés sur les différentes présentations orales que j'ai effectué durant ces trois ans ; François WAHL et Pascal DUCHESNE pour leur expertise sur le krigeage. Mickaël RIVALLAN et Anne-Agathe QUOINEAUD pour leur expertise sur la Résonance Magnétique Nucléaire.

Je remercie tous les membres du Jury qui ont accepté de juger ces travaux de thèse.

Remerciements tout particuliers à ma famille qui m'a soutenu durant ces trois ans.

Enfin, je conclurai en remerciant Dieu de m'avoir orienté vers cette entreprise et de m'avoir ouvert les portes d'un monde que je n'aurais jamais pensé côtoyer. A lui revienne toute la Gloire. Amen.

Sommaire

Listes des abréviations

ACP : Analyse en composantes principales

API.Gravity : Inverse de la gravité spécifique

ATR : *Attenuated Total Reflectance*

C_a : Proportion de carbone aromatique

C_i : Hydrocarbure possédant i atomes de carbone

C_{i+} : Hydrocarbure possédant au moins i atomes de carbone

C_{ip} : Proportion de carbones isoparaffiniques

C_n : Proportion de carbones naphténiques

C_{np} : Proportion de carbones n-paraffiniques

C_p : Proportion de carbones paraffiniques

d : densité par rapport à l'eau

$d_{t_1}^{t_2}$: densité d'un liquide à la température t_2°C par rapport à l'eau à la température t_1°C

DS : Distillation simulée

DSV : Distillat sous vide ou VGO : *vacuum gas oil*

EHD : Expérimentations hauts débits

FBF : *Final Boiling Point* ou point d'ébullition final

FID : *Free Induction Decay*

GC : Chromatographie en phase gazeuse monodimensionnelle

GC×GC : Chromatographie en phase gazeuse bidimensionnelle

GPL : Gaz de pétrole liquéfié

HCK : Hydrocraquage

HDA : Hydrogénation des aromatiques

HDT : Hydrotraitement

IBF : *Initial Boiling Point* ou point d'ébullition initial

IC : Intervalle de confiance

IR : Indice de réfraction

IR : Infrarouge

K_w : Facteur de caractérisation de Watson

LC : Chromatographie en phase liquide

Liqtot : Effluent total du réacteur

M : Masse molaire

MAD : *Mean absolute deviation* ou erreur absolue moyenne

MeABP : *Mean Average Boiling Point* ou point d'ébullition moyen

MRD : *Mean relative deviation* ou erreur relative absolue moyenne

MS : *Mass spectrometry* ou Spectrométrie de masse

PE : Point d'écoulement

PLS : Partial least squares

PT : Point de trouble

Ra : Nombre moyen de cycles aromatiques par molécules

RMN : Résonance Magnétique Nucléaire

RMSE : *Root mean square error* ou erreur quadratique moyenne

RMSEC : *Root mean square error of calibration* ou erreur quadratique moyenne de calibration

RMSECV : *Root mean square error of cross-validation* ou erreur quadratique moyenne de validation croisée

RMSEP : *Root mean square error of prediction* ou erreur quadratique moyenne de prédiction

Spgr : Specific gravity

TLF : Température limite de filtrabilité

T_M : Température moyenne

T_{MP} : Température moyenne pondérée

T_i : Température à laquelle i% de la masse d'un produit a été distillé

UCO : Unconverted oil ou fraction d'huile non convertie

VI : Indice de viscosité

VI.chg : Indice de viscosité de la charge

VVH : Vitesse volumétrique horaire

X_{370} : Taux de conversion de la fraction huile

Liste des tableaux

Listes des illustrations

Liste des annexes

Introduction

L'industrie pétrolière est de plus en plus confrontée à une dégradation de la qualité des pétroles bruts [1]. En effet, la part de distillats moyens contenue dans les pétroles bruts à partir desquels les carburants (essence, kérosène, gazole,…) sont obtenus est de plus en plus faible [2]. *A contrario*, la part de fractions lourdes de ces pétroles, valorisable pour l'essentiel que comme fiouls lourds et bitumes est de plus en plus prépondérante. La conversion de ces coupes lourdes en coupes carburants est donc devenue un enjeu majeur de l'industrie du raffinage pétrolier afin de répondre à la demande croissante de l'industrie du transport [2]. Les pétroles bruts sont des mélanges complexes de composés organiques, majoritairement des hydrocarbures [3]. Par distillation atmosphérique [4] et différentes étapes de raffinage, plusieurs produits commercialisables sont obtenus. Il peut s'agir de gaz, d'essence, de kérosène, de gazole ou de lubrifiants, par intervalles d'ébullition croissants, des coupes légères aux coupes lourdes. Au-delà des aspects quantitatifs, les produits pétroliers sont soumis à des spécifications fixées par les normes européennes ou internationales. Ces spécifications concernent différentes propriétés notamment liées à la combustion, à l'écoulement, à la tenue à froid ou encore à des contraintes environnementales [3,5].

Dans ce contexte, l'hydrocraquage est un procédé de choix qui, à partir de distillats sous vide du pétrole brut, permet d'obtenir des produits pétroliers de qualité [2,6]. Atteindre les spécifications requiert des conditions opératoires optimales, **ainsi que le développement de catalyseurs performants [2,6–10].** La mise en place et l'optimisation du procédé d'hydrocraquage nécessite donc de réaliser des expérimentations (essais pilotes). De plus, la mesure des propriétés des produits nécessite souvent un volume d'échantillon significatif (une centaine de ml) et un temps d'analyse élevé, ce qui entraine **un coût financier important.**

Pour réduire les coûts liés à l'optimisation des procédés d'hydrocraquage un simulateur (dénommé simulateur IFPEN), contenant des modèles de prédiction de propriétés de produits a ainsi été développé au sein d'IFPEN [7,11–13]. Ces modèles permettent d'estimer les propriétés d'intérêt à partir de données analytiques de la charge, des effluents et de données expérimentales complémentaires. L'avantage de la simulation est double : **gain de temps considérable et réduction du nombre d'expérimentations**. Etant donnés les enjeux, la qualité des prévisions fournies par les modèles est capitale. A ce jour, certaines propriétés **sont mal prédites** pour deux raisons : **une mauvaise reproductibilité des méthodes de mesures** et un manque de compréhension qui se traduit par **une remise en question de la pertinence des descripteurs.**

Toujours dans cet objectif de réduction des coûts, une alternative complémentaire est la mise en place d'unités d'expérimentation haut débit (EHD) qui permettent un *screening* rapide de plusieurs catalyseurs en parallèle pour accélérer leur mise au point. La quantité d'effluent total fournie par ce type d'unité (quelques ml) est cependant insuffisante pour effectuer l'étape de distillation préalable à

l'obtention des coupes pétrolières. Il est donc impossible d'accéder à ces coupes et de s'assurer de leur qualité [14,15].

Les travaux de cette thèse doivent permettre de répondre aux trois problématiques suivantes :

1. **Comment accéder aux propriétés des produits lorsque les expérimentations sont faites sur des unités EHD ?**
2. **Peut-on améliorer la compréhension moléculaire des propriétés d'intérêt ?**
3. **Comment améliorer les performances prédictives du simulateur pour les propriétés d'intérêt ?**

Ces travaux concernent particulièrement deux propriétés des produits pétroliers, actuellement mal modélisées par le simulateur IFPEN qui sont le point de trouble (PT) de la coupe gazole et l'indice de viscosité (VI) de la coupe huile. Le PT désigne la température à laquelle un produit commence à cristalliser tandis que le VI mesure la dépendance de la viscosité d'un produit en fonction de la température. La mesure ou l'estimation de ces propriétés est essentielle pour s'assurer de la possibilité de l'utilisation des produits dans les moteurs [16].

Compréhension moléculaire et modélisation de propriétés sont liées : (1) mieux comprendre les facteurs qui influent sur une propriété donnée permet d'améliorer la pertinence des descripteurs utilisés pour sa modélisation ; (2) l'étude *a posteriori* d'un modèle prédictif peut fournir des informations essentielles sur l'impact des facteurs qui régissent cette propriété.

La modélisation des propriétés d'intérêt est le cœur des travaux de cette thèse. Elle repose sur une méthodologie basée sur différentes étapes qui sont décrites ci-dessous.

La première étape consiste à collecter des échantillons pour les coupes pétrolières d'intérêt.

La seconde étape est la caractérisation des échantillons collectés. Il existe des techniques de mesure de propriétés physico-chimiques dites globales (densité, indice de réfraction, courbe de distillation, *etc.*), généralement normalisées et qui sont couramment mises en œuvre par les raffineurs [4,17–27]. D'autres techniques dites de caractérisation moléculaire sont de plus en plus utilisées. Elles ont pour objectif de fournir une cartographie des différentes espèces moléculaires présentes dans les mélanges complexes. Parmi ces techniques on retrouve notamment les techniques spectroscopiques (Infrarouge [28–31], résonance magnétique nucléaire [32–34], *etc.*) et les techniques chromatographiques (chromatographie en phase gazeuse mono et bidimensionnelle [35,36], chromatographie liquide [37–39], *etc.*). Outre la quantité d'information qu'elles fournissent, ces techniques présentent l'avantage de ne requérir qu'une très faible quantité de produits (généralement quelques µL) [35].

La troisième étape consiste en la validation des données collectées. En effet, la présence dans la base de données de valeurs aberrantes aussi appelées « *outliers* », peut significativement dégrader la qualité d'un modèle prédictif et fausser son interprétation [40,41]. Un objectif connexe de cette thèse

est d'effectuer un état de l'art des méthodes de détection de points aberrants pour déterminer la (les) mieux adaptée(s) à la modélisation de propriétés physico-chimiques dans les produits pétroliers.

La quatrième étape concerne le choix de descripteurs pertinents. En complément de l'identification des principaux facteurs qui influent sur la propriété à modéliser, les méthodes statistiques de sélection de variable sont des outils essentiels pour le développement de modèles performants [42–49].

La cinquième étape est le développement du modèle proprement dit.

La sixième étape consiste en l'analyse *a posteriori* du modèle obtenu (diagnostic des poids attribués aux variables, analyse de sensibilité [50,51], *etc.*).

Pour améliorer la compréhension moléculaire des propriétés d'intérêt nous proposons une approche basée sur la modélisation prédictive des propriétés à partir de données issues de la caractérisation moléculaire des échantillons de type gazole et huile. L'objectif est d'analyser les modèles développés, notamment les modèles de régression PLS (*Partial Least Squares*) [52,53] et de régression PLS parcimonieuse (*sparse* PLS) [54,55], pour identifier les principaux marqueurs moléculaires qui influent sur la propriété modélisée.

Pour améliorer la qualité de la prédiction des propriétés d'intérêt dans le simulateur IFPEN, il faut d'une part effectuer un choix le plus pertinent possible des descripteurs et d'autre part introduire des modèles prédictifs adaptés aux données dont on dispose. Ici, l'utilisation des données de caractérisation moléculaire des échantillons dans le simulateur IFPEN est inenvisageable car le raffineur n'y a pas accès. De ce fait, seules des propriétés globales des coupes mises en jeu lors du procédé peuvent être utilisées comme descripteurs.

D'autre part, les modèles de régression (linéaire et non linéaire) utilisés [46,56,57] actuellement, qui présentent l'avantage d'être facilement interprétables et relativement simple à optimiser deviennent très difficiles à construire voire inadaptés quand le nombre de descripteurs augmente. En effet, ce type de modèle repose sur une forme analytique explicite permettant de relier la propriété aux descripteurs. Nous proposons donc l'utilisation de méthodes locales basées sur le principe d'interpolation [58]. Introduites pour résoudre des problèmes de géophysique, ces méthodes qu'elles soient déterministes (de type interpolation par *splines* [59,60]) ou stochastiques (krigeage [61,62]) n'ont à notre connaissance jamais été utilisées pour la prédiction de propriétés physico-chimiques des produits pétroliers. De par leur construction, les méthodes stochastiques permettent d'associer à chaque prévision une estimation de son incertitude. **Ce point constitue un atout majeur dans la prédiction de propriétés physico-chimiques de produits pétroliers.**

Le manuscrit est structuré comme suit : la première partie présente tout d'abord le contexte de la thèse, centrée sur le procédé d'hydrocraquage ; un état de l'art sur la compréhension moléculaire et la prédiction des propriétés d'intérêts y est proposé. La seconde partie fournit : (1) une présentation théorique des différentes techniques analytiques qui ont été utilisées pour la caractérisation des coupes pétrolières mises en jeu ; (2) une description des méthodes de statistique prédictives mises en œuvre

pour exploiter les données ; (3) les différentes bases de données utilisées pour ces travaux. La troisième et dernière partie, présente et discute les résultats obtenus. Un chapitre consacré aux méthodes de détection de points aberrants et à leur application à la modélisation du VI de la coupe huile est également présenté en **Annexe M**.

Partie 1 : Contexte et état de l'Art

L'objectif de cette première partie est double. Il consiste d'abord à initier le lecteur au domaine du raffinage du pétrole brut pour l'obtention de produits pétroliers et à introduire la terminologie propre à ce domaine, puis à présenter globalement les méthodes d'analyse statistique de données et leurs applications à la compréhension et à la modélisation de propriétés importantes des produits pétroliers.

Dans le Chapitre 1, nous détaillerons dans un premier temps la composition des pétroles bruts en insistant sur la complexité du système et les différentes coupes pétrolières très importantes pour l'industrie du raffinage. Nous présenterons ensuite le procédé d'hydrocraquage des distillats sous vide (HCK), en insistant sur les deux coupes qui vont mobiliser notre attention : la coupe gazole et la coupe huile.

Dans le Chapitre 2, une étude bibliographique sur la compréhension moléculaire et la modélisation des propriétés d'intérêt de ces coupes sera proposée, de même qu'une présentation globale des notions de statistiques prédictives. Enfin, un bilan sur l'état de l'art sera effectué et la méthodologie mise en place pour atteindre les objectifs de cette thèse sera clairement décrite.

Chapitre 1. Pétrole brut et Raffinage

L'objet de ce chapitre est de présenter les généralités sur le pétrole brut et les produits pétroliers issus de son raffinage, ainsi que les principaux fondements du procédé d'hydrocraquage.

1.1 Composition des pétroles bruts

Le pétrole est un mélange complexe majoritairement constitué d'hydrocarbures (93 à 99% m/m) mais également de composés organiques soufrés (0,01 à 6% m/m), azotés (0,05 à 0,5 % m/m), oxygénés (0,1 à 0,5 % m/m) et de certains métaux (0,005 à 0,15 % m/m) tels que le nickel et le vanadium [3]. Il est composé d'un continuum de molécules hydrocarbonées pouvant comporter de quelques unités à plus d'une centaine d'atomes de carbones.

La Figure 1 illustre la composition des pétroles bruts en fonction de la température d'ébullition et du nombre d'atomes de carbone. Les différentes familles d'hydrocarbures présentes dans les pétroles (paraffines, naphtènes, aromatiques) y sont également mentionnées, de même que la constitution de certaines coupes pétrolières telle que l'essence (*naphtha*), les distillats moyens (kérosène et gazole) ou encore les distillats sous vide (*VGO*) et les résidus sous vide (*Vacuum Residue*). Ce schéma illustre le fait que les matrices pétrolières sont très complexes, le nombre d'isomères augmentant significativement avec le nombre de carbones.

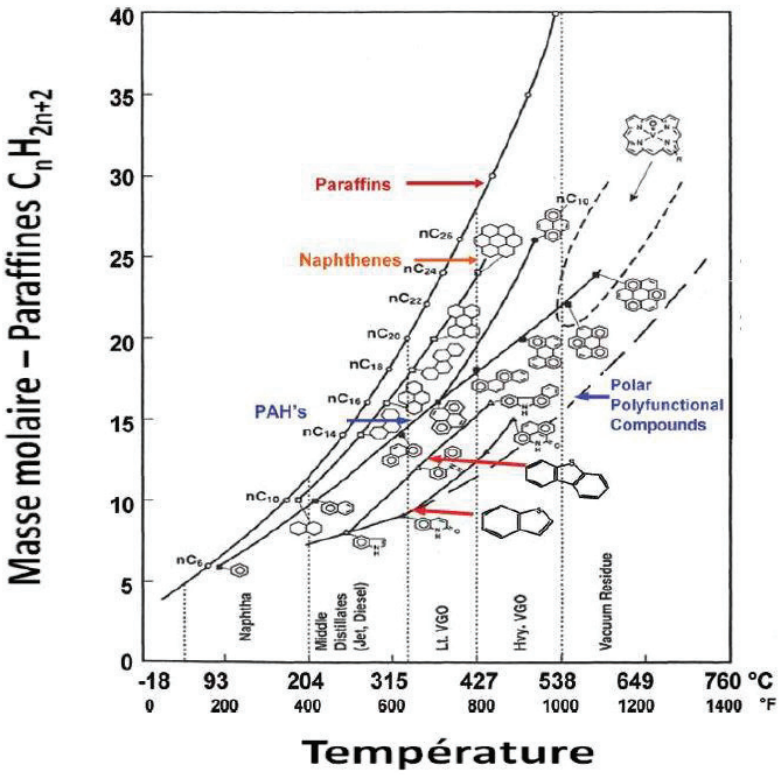


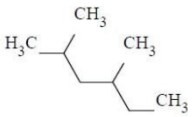
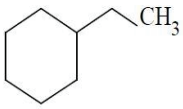
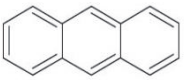
Figure 1 : Composition d'un pétrole brut et relation point d'ébullition / masse molaire / structure [63]

1.1.1 Les familles d'hydrocarbures présents dans le pétrole brut

Les produits pétroliers qui sont issus du fractionnement du pétrole brut sont majoritairement constitués d'hydrocarbures de différentes familles chimiques. On distingue 4 principales familles en fonction du degré d'insaturations lié à la structure des molécules [3] :

1. les **normales paraffines (n-paraffines)** ou alcanes linéaires qui possèdent une chaîne droite saturée,
2. les **iso-paraffines** ou alcanes linéaires saturés ramifiés
3. les **naphtènes** qui contiennent au moins une chaîne cyclique carbonée saturée de 5 ou 6 atomes ;
4. les **hydrocarbures aromatiques** qui sont des composés cycliques polyinsaturés présents en forte quantité dans les coupes les plus lourdes. Ils peuvent contenir un ou plusieurs cycles aromatiques et / ou naphténiques et / ou des chaînes ramifiées.

Tableau 1 : Structure de composés hydrocarbonés présents dans les fractions pétrolières [30]

Familles	n-Paraffines	iso-Paraffines	Naphtènes	Aromatiques
Formules brutes	C_nH_{2n+2}	C_nH_{2n+2}	C_nH_{2n}	C_nH_{2n-8k}
Exemples				

$n \rightarrow$ nombre d'atomes de carbone ; $k \rightarrow$ nombre de cycle benzénique

1.1.2 Les composés soufrés, azotés, oxygénés et organométalliques

Les autres types de composés (soufrés, azotés, *etc.*) présents dans les pétroles bruts doivent être éliminés au cours des différents procédés de raffinage du fait de leur toxicité et de leur effet négatif sur les catalyseurs ou sur la qualité des produits. Des exemples de molécules appartenant à ces différentes familles chimiques sont donnés en Annexe A.

Les pétroles bruts et certaines fractions pétrolières lourdes contiennent notamment une proportion notable de soufre. Le soufre, élément chimique divalent, est associé à l'hydrogène et au carbone dans 4 types de composés principaux [3] :

- les mercaptans dans lesquels l'hydrogène lié au soufre a un caractère acide,
- les sulfures où le soufre est intercalé dans une chaîne saturée,
- les disulfures, de formule générale R – S – S – R',
- les thiophènes, où le soufre est inséré dans un cycle aromatique.

D'une manière générale, la teneur en azote des bruts et des fractions pétrolières est bien moindre que la teneur en soufre. Ces composés se distinguent suivant leur caractère basique ou neutre et sont connus pour empoisonner les catalyseurs acides [3].

Les composés oxygénés sont constitués de carbone, d'hydrogène et d'oxygène. Les plus courants sont les acides naphténiques, qui sont des acides organiques présents dans certains bruts, et les dérivés du phénol.

Les métaux (essentiellement nickel et vanadium), sont présents dans les bruts au sein des fractions les plus lourdes en faible quantité. Ils sont combinés au sein de très grosses molécules renfermant en général tous les éléments déjà cités : carbone, hydrogène, soufre, azote, oxygène. Ils sont également présents dans certains composés plus petits de la famille des porphyrines où le motif de base est constitué par un ensemble de quatre cycles pyrroliques, le métal étant au centre de cet ensemble sous la forme Ni^{++} ou VO^{+} [3]. Afin de transformer le pétrole brut en produits valorisables différentes étapes de raffinage sont mises en œuvre.

1.2 Raffinage et coupes pétrolières

1.2.1 Distillation du pétrole

La distillation est une étape préliminaire au raffinage des produits pétroliers [3,4]. Elle permet de fractionner le pétrole brut afin d'obtenir différentes coupes pétrolières en fonction de la température d'ébullition. La Figure 2 représente un schéma simplifié de la distillation du pétrole brut. Les points de coupes qui apparaissent sur ce schéma sont les standards utilisés à IFPEN [35]. Chacune des coupes pétrolières correspond à un intervalle de volatilité, que l'on peut caractériser par une gamme de températures d'ébullition, ou par le nombre d'atomes de carbone des hydrocarbures. La distillation atmosphérique permet de séparer les coupes gaz (<50°C), la coupe essence (50 – 150°C), la coupe kérosène (150 – 250°C) et la coupe gazole (250 – 370°C). La partie du produit qui n'a pas été distillée lors de cette opération est appelée le Résidu Atmosphérique (RA). Elle est composée de molécules dont le point d'ébullition est supérieur à 370°C. Les Distillats Sous Vide (DSV) correspondent aux fractions les plus légères de cette coupe résiduelle séparée par distillation sous vide [3]. Ils sont constitués de composés qui ont entre 20 et 50 atomes de carbone et une température d'ébullition comprise entre 370 et 550°C environ. Les coupes DSV se distinguent des fractions résiduelles non distillables (Résidu sous vide) notamment par le fait qu'elles contiennent peu d'asphaltènes et de métaux. De ce fait, ces coupes se prêtent plus facilement aux opérations de conversion catalytique.

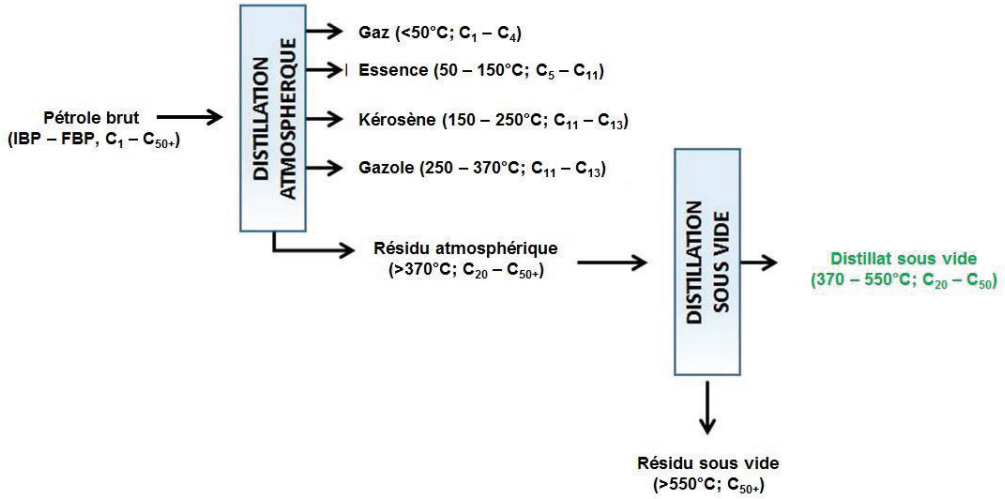


Figure 2 : Distillation atmosphérique et sous vide d'un pétrole brut et exemples de coupes standard IFPEN ; IBP → *Initial Boiling Point* ; FBP → *Final Boiling Point* [3]

La conversion des DSV fait essentiellement appel au craquage catalytique (ou *Fluid Catalytic Cracking*) surtout producteur d'essences et à l'hydrocraquage (HCK) qui permet, lui, de générer surtout des coupes kérosène et gazole de très bonne qualité [2].

1.2.2 Le procédé d'hydrocraquage

Malgré des coûts d'investissement très élevés (en raison de conditions opératoires sévères), l'HCK est devenu un procédé très utilisé dans les raffineries. Cela s'explique par l'accroissement de la demande en carburants (essence, kérosène et gazole). Le procédé existe sous différentes formes suivant l'intensité de craquage souhaitée : à basse pression [64] (aussi connu sous le nom de *mild hydrocracking*) et à haute pression, l'objectif étant toujours de convertir la plus grande partie de la charge en produits plus légers [2,7,9].

La Figure 3 représente un schéma simplifié du procédé d'HCK des DSV. Les réactions se déroulent sous haute pression d'hydrogène (50 à 180 bars), à haute température (380 à 430°C) et en présence de catalyseurs. Le procédé d'HCK est constitué de deux étapes : une étape d'hydrotraitement (réacteur HDT, Figure 3) qui vise à éliminer le soufre (hydrodésulfuration) et l'azote (hydrodésazotation). L'élimination de l'azote est primordiale en raison de l'effet néfaste de certains composés azotés très basiques sur le catalyseur acide d'HCK. Sa concentration dans l'effluent hydrotraité doit être faible (comprise entre 0 et 300 ppm). La seconde étape est l'HCK (réacteur HCK, Figure 3) à proprement parler, qui a pour fonction de craquer les molécules lourdes présentes dans l'effluent hydrotraité en fractions plus légères. Les molécules sont isomérisées puis craquées. Les réactions d'HCK sont réalisées sur des catalyseurs spécifiques possédant les propriétés acides requises pour les réactions de craquage [7]. L'effluent hydrocraqué (liqtot) est récupéré en sortie du réacteur

HCK (Figure 3), puis distillé dans les conditions atmosphériques pour obtenir les différents produits (gaz, essence, kérosène et gazole). Le rendement des différentes coupes dépend de la conversion appliquée, du catalyseur et des conditions opératoires. Ces coupes doivent répondre à des spécifications liées à leur volatilité, leur combustion, leur tenue à froid, leur écoulement, la pollution atmosphérique, *etc.* pour pouvoir être vendues sous forme de carburant [16]. Ces propriétés sont mesurées sur les coupes physiques par des méthodes normalisées (ASTM ou NF EN ISO, Figure 3).

Le réacteur HCK produit en outre, avec un rendement dépendant de la conversion, une fraction lourde non convertie (Figure 3) appelée *unconverted oil* (UCO) 370^{+} [35]. Cette dernière est déparaffinée à l'aide d'un solvant (Méthode IFPEN) [14] ou d'un catalyseur [65] pour l'obtention d'une huile de base. Ce produit est lui aussi soumis à différentes spécifications relatives notamment à son VI qui est une des plus importantes [3].

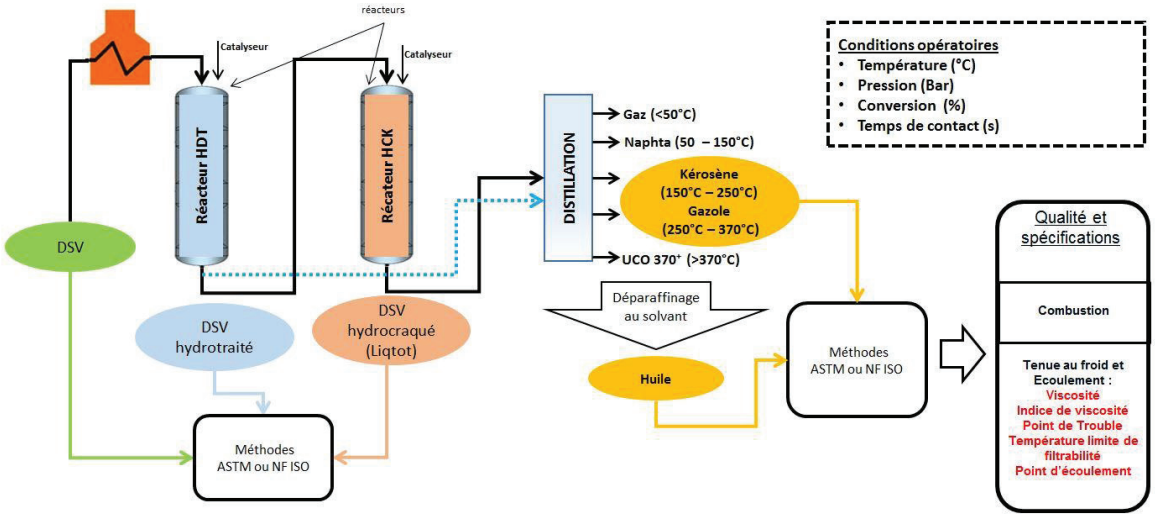


Figure 3 : Schéma généralisé du procédé d'hydrocraquage

1.2.3 Les principales cibles du procédé d'hydrocraquage

Comme nous l'avons mentionné auparavant, le PT de la coupe gazole et le VI de la coupe huile sont des propriétés importantes mais qui sont pour le moment mal comprises et donc mal prédites. C'est pourquoi dans ce manuscrit, nous nous concentrerons uniquement sur les coupes gazole et huile.

1.2.3.1 La coupe gazole

La coupe gazole est un mélange d'hydrocarbures purs dont les points de coupes (standard IFPEN) sont compris entre 250 et 370°C. Pour cette coupe, la longueur des chaînes carbonées varie entre environ 13 et 25 atomes de carbone. La qualité d'un gazole se caractérise principalement par les spécifications des carburants présentées dans le Tableau 2 [16].

Tableau 2 : Spécifications relatives aux gazoles moteur en France [3]

Caractéristiques	Spécifications	Méthodes ASTM/NF/EN
Masse volumique à 15°C (kg/m^3)	Entre 820 et 845	EN ISO 12185
Distillation en %vol	<10% à 180°C >95% à 340°C	EN ISO 3924
Viscosité à 40°C (mm^2/s)	$1,5 \leq \nu < 4$	EN ISO 3104
Teneur en soufre (mg/kg)	≤10	EN ISO 20846
Stabilité à l'oxydation (g/m^3)	≤25	EN ISO 12205
Indice de cétane mesuré	>51	EN ISO 5165
Indice de cétane calculé	>46	EN ISO 4264
Point de trouble (°C)	≤+5°C d'Avril à Septembre ≤-10°C d'Octobre à Mars	ASTM D2500 / EN 23015
Température limite de filtrabilité (°C)	≤0°C d'Avril à Septembre ≤-20°C d'Octobre à Mars	ASTM D6371 / EN 116
Point d'écoulement (°C)	Pas de spécification	ASTM D97 / T60-105

1.2.3.2 La coupe huile

La coupe huile est un mélange d'hydrocarbures dont la température d'ébullition est généralement supérieure à 370°C. La qualité d'une huile est caractérisée par son VI, son point d'écoulement (PE), ainsi que par sa stabilité thermique à l'oxydation [66].

L'American Petroleum Institute (API) classe les huiles de base selon les critères définis dans le Tableau 3 [66]. Parmi les cinq catégories proposées (I-V), trois se réfèrent à des huiles paraffiniques (I-III) et une à des huiles naphténiques (IV). Les huiles paraffiniques sont quant à elles classées selon **leur VI qui caractérise l'impact de la température sur la variation de leur viscosité**, et selon la teneur en aromatiques, en saturés et leur concentration en soufre. Les catégories II^+ et III^+ ne sont pas des catégories officielles mais sont employées au niveau du marketing des huiles et ont donc été indiquées dans ce tableau. Dans le cas du procédé d'HCK, les huiles ciblées sont généralement de groupe III (VI>120).

Tableau 3 : Catégories des différentes huiles de base [66]

Groupes API	% saturés	VI	%m/m en Soufre	Remarques
I	<90	>80 à <120	>0,03	Raffinage au solvant (conventionnel)
II	>90	>80 à <120	<0,03	Nécessité d'un *hydroprocessing*
II+	>90	>110 à <119	<0,03	
III	>90	>120	<0,03	Sévère *hydroprocessing* : HCK, GTL (gas to liquid)
III+	>90	>120 à <150	<0,03	
IV	Polyalphaoléfines (PAO)			PAO ou huile de synthèse
V	Toute huile non inclue dans les groupes I-IV (esters, huiles pâles)			Inclus les autres synthèses

Assurer la qualité des produits (ici gazoles et huiles) pour répondre aux spécifications, constitue l'objectif principal du raffineur. Pour cela, il est indispensable de mieux comprendre les propriétés mises en cause.

1.2.4 Les différentes réactions en HCK : hydrotraitement

Les réactions qui ont lieu au cours de l'HDT permettent de transformer des molécules organiques azotées et soufrées en d'autres composés plus légers et moins néfastes pour le catalyseur comme l'ammoniac (NH_3) et le sulfure d'hydrogène (H_2S) [10].

Les réactions d'HDT sont généralement classées en trois grandes familles [10] :

1. l'hydrodésulfuration (HDS)
2. l'hydrodésazotation (HDN)
3. l'hydrogénation des aromatiques (HDA)

Les réactions d'HDS ont pour but d'éliminer le soufre organique contenu dans la charge et ainsi de permettre le respect des spécifications imposées sur les produits pétroliers [16]. La totalité des composés soufrés est traitée dans les conditions habituelles d'HDT. Les réactions d'HDN sont particulièrement importantes compte tenu de l'impact des composés organiques azotés sur les catalyseurs d'HCK. En

effet, ces derniers empoisonnent fortement les sites acides des zéolithes et inhibent les réactions d'HDT et les réactions d'HCK. Cette réaction permet de traiter tous les composés azotés présents dans la charge. L'HDA est une réaction qui précède le craquage des hydrocarbures cycliques. En effet, le craquage direct d'un composé aromatique n'est pas possible. Ces réactions d'hydrogénation sont réversibles et leur vitesse dépend fortement du catalyseur employé. Des exemples de réactions d'HCK sont donnés en Annexe B (Tableau A. 3).

1.2.5 Les différentes réactions en HCK : hydrocraquage

Le procédé d'HCK est catalytique [2,10]. Il utilise un catalyseur bifonctionnel ayant à la fois [10] :

- une fonction hydrogénante / déshydrogénante (H/DH) de type métallique comme des métaux nobles (platine, palladium) ou non nobles du groupe VIA (molybdène, tungstène) et du groupe VIIIA (cobalt ou nickel).
- une fonction acide Brönsted de type silice-alumine (amorphe) ou zéolithe, c'est-à-dire des silice-alumines cristallisées (souvent US-Y pour Ultra Stabilized Y)

En présence d'hydrogène, le rôle de la fonction H/DH est de déshydrogéner les alcanes et d'hydrogéner les alcènes [10]. Les métaux nobles nécessitent des charges avec une teneur faible en composés organiques soufrés et azotés, contrairement aux métaux utilisés sous forme soufrée. La fonction Brönsted a pour rôle quant à elle d'isomériser et de craquer les carbocations [10]. L'équilibre entre la réactivité de ces deux fonctions permet de gérer la sélectivité. Des exemples de réactions d'HCK sont donnés en Annexe B (Tableau A. 5).

Chapitre 2. Compréhension moléculaire et prédiction des propriétés produits

Mieux comprendre une propriété d'intérêt c'est identifier les principaux marqueurs moléculaires (nature et structure des hydrocarbures présents dans les produits) qui ont une influence sur sa variation et définir le sens de cet impact. Par ailleurs, la modélisation de ces propriétés d'intérêt est un challenge important pour les raffineurs étant donnés les enjeux économiques et financiers.

Dans cette optique, de nombreuses études ont été effectuées autour du PT de la coupe gazole [67–70] et du VI de la coupe huile [66,71–73]. La majeure partie de ces études font appel d'une part à des techniques de caractérisation des produits pétroliers et d'autre part à des méthodes d'analyse de données multidimensionnelles et multivariées. Une difficulté majeure de l'industrie pétrolière réside dans le fait que la quantité d'échantillons analysés pour générer les données est relativement faible. De ce fait, l'utilisation de certaines méthodes statistiques peut être remise en question.

Dans ce chapitre, nous revenons tout d'abord sur les généralités autour des méthodes statistiques prédictives couramment utilisées pour la modélisation de propriétés des produits pétroliers. Ensuite, les différentes études qui ont été recensées sur la compréhension moléculaire et la prédiction du PT de la coupe gazole et du VI de la coupe huile sont présentées. Pour chaque coupe, les généralités sur les propriétés d'intérêt sont rappelées, puis une discussion sur la pertinence et la fiabilité des différents travaux est proposée.

2.1 Méthodes statistiques prédictives

Plusieurs méthodes d'analyse de données basées sur des notions de statistique prévisionnelle ont été développées au cours des dernières décennies [57,74–76]. Parmi ces méthodes statistiques prédictives, les méthodes dites d'apprentissage supervisé sont les plus utilisées dans le cadre de la modélisation de propriétés physico-chimiques de produits pétroliers. Nous rappelons ci-dessous la méthodologie de l'apprentissage statistique supervisé.

2.1.1 Méthodologie de l'apprentissage statistique supervisé

L'apprentissage statistique concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou problématiques par des moyens algorithmiques plus classiques [57]. Elle traite du problème de la recherche d'une fonction prédictive basée sur des données. L'objectif général de l'apprentissage statistique est la modélisation qui peut se préciser en sous-objectifs à définir clairement préalablement à une étude car ceux-ci conditionnent en grande partie les méthodes qui pourront être mises en œuvre [74] :

- modéliser pour explorer, représenter, décrire les variables et leurs liaisons ;

- modéliser pour expliquer l'influence d'une variable ou facteur dans un modèle supposé connu *a priori* ;
- modéliser pour prévoir et sélectionner un meilleur ensemble de prédicteurs ;
- modéliser juste pour prévoir.

Il existe deux grands groupes de méthodes d'apprentissage selon la présence ou non d'une variable Y à expliquer [57,74] :

1. l'apprentissage supervisé qui consiste à trouver une fonction f susceptible d'approcher « au mieux » Y
2. l'apprentissage non-supervisé où l'objectif et généralement la recherche d'une typologie ou taxinomie des observations.

L'apprentissage statistique supervisé consiste généralement en six étapes décrites ci-dessous [57,74] :

1. Constitution de la base de données (collecte et analyse d'échantillons)
2. Exploration des données pour la détection de valeurs aberrantes ou atypiques, d'incohérences, pour l'étude des distributions, des structures de corrélation, recherche de typologies, transformations de données, *etc.*
3. Partition des données en trois bases (base d'apprentissage, base de validation et base de test)
4. Etape dite d'apprentissage ou de calibration qui consiste à estimer le modèle pour une valeur donnée d'un paramètre (ou de plusieurs) de *complexité* : nombre de variables, paramètres de seuillage, *etc*.
5. Etape de validation (si nécessaire) qui consiste à fixer au mieux ce(s) paramètre(s) suivant un critère prédéfini. Les étapes 4 et 5 sont répétées autant de fois que nécessaire.
6. Etape de test qui permet de juger de la qualité du modèle obtenu à l'issue des étapes 1 et 2.

Suivant le modèle considéré, l'étape de validation n'est pas toujours requise. Par ailleurs, lorsqu'on dispose d'une base de données limitée les étapes d'apprentissage et de validation peuvent être confondues. Dans ce cas, on utilise des méthodes dites de « validation croisée ». Ces différentes étapes de l'apprentissage statistique sont indispensables pour s'assurer de la qualité d'un modèle prédictif.

Les méthodes d'apprentissage statistique supervisé peuvent être classées en deux groupes [57] :

1. Les méthodes d'apprentissage paramétrique et semi-paramétrique
2. Les méthodes d'apprentissage non paramétrique

Nous nous limitons ici au cas des modèles de régression, c'est-à-dire que la variable à modéliser est quantitative.

2.1.1.1 Méthodes d'apprentissage paramétrique

Le principe des méthodes d'apprentissage paramétrique est d'approcher au mieux la variable réponse y par une fonction mathématique dont la forme analytique est prédéfinie. Notons f cette fonction. On appelle modèle de régression paramétrique toute équation de la forme [57] :

$$y = f(x, \theta) + \varepsilon \qquad \textbf{(Eq.1. 1)}$$

où θ désigne l'ensemble des paramètres intrinsèques associés à l'expression analytique de f et ε représente l'erreur de modélisation incluant le bruit sur la mesure de référence. Il existe deux types de modèle de régression suivant la forme de la fonction f un modèle est dit linéaire (au sens général) si f peut s'écrire comme une combinaison linéaire des composantes de θ (fonctions affines, polynômes, *etc.*) ; sinon on parle de modèles non linéaires (cinétiques, thermodynamiques, *etc.*) [57]. Ces méthodes, particulièrement simples à implémenter ont toutefois certaines limites : d'abord, la forme analytique de la fonction de modélisation influe fortement sur la qualité du modèle ; ensuite, les méthodes classiques d'estimation des paramètres ne sont en général valables que lorsque le nombre d'observations est supérieur au nombre de variables explicatives (condition qui n'est pas toujours vérifiée en chimiométrie).

2.1.1.2 Méthodes de régression multivariée ou chimiométriques

Les méthodes de régression multivariées permettent de surmonter les difficultés liées à la dimension de l'espace d'étude. Elles sont basées sur la notion de réduction d'espace qui consiste à substituer les variables explicatives initiales à de nouveaux facteurs en leur appliquant une transformation de sorte que le nombre r de nouveaux facteurs soit inférieur au nombre d'observations n. La construction des facteurs de substitution se fait de manière itérative. L'objectif est de conserver le maximum d'information essentielle (au sens statistique). Les méthodes les plus souvent utilisées sont les méthodes de régression sur composantes orthogonales qui consistent à substituer aux variables initiales des facteurs deux à deux orthogonaux entre eux et qui sont des combinaisons linéaires des variables initiales. C'est le cas notamment de la PCR (*Principal Components Regression*) [57,74] et de la régression PLS (*Partial Least Squares*) [52,53].

2.1.1.3 Méthodes d'apprentissage non paramétrique

En l'absence de toute hypothèse sur la fonction de modélisation (contrairement au cas de l'apprentissage paramétrique) on parle de méthodes d'apprentissage non paramétrique. Certaines d'entre elles consistent à estimer la fonction de modélisation par une somme de fonctions élémentaires qui constituent une base de l'espace d'étude (estimation par des polynômes par morceaux, estimation sur des bases de *splines*, estimation par noyaux, estimation par polynômes locaux, estimation par projection sur des bases orthonormées, *etc.*) [57,74]. D'autres méthodes d'apprentissage non paramétrique plus sophistiquées sont de plus en plus préconisées pour la modélisation de propriétés complexes. C'est le cas des réseaux de neurones [77] ou des machines à vecteurs supports [78,79].

L'apprentissage non paramétrique offre plus de flexibilité que l'apprentissage paramétrique puisqu'il permet une adaptation automatique à des situations diverses (linéarité, non linéarité, irrégularité, *etc*). Les méthodes d'apprentissage non paramétrique sont cependant plus longues à implémenter et dans le cas de l'apprentissage automatique, elles requièrent une base de données

significative pour assurer la qualité du modèle. Les modèles obtenus sont par ailleurs difficiles à interpréter, ce qui explique sans doute pourquoi elles sont encore relativement peu utilisées en chimiométrie.

2.1.2 Qualité d'un modèle prédictif

Les performances d'un modèle prédictif s'évaluent par la qualité de ses prévisions. Cette dernière est caractérisée à la fois par la précision, c'est-à-dire la capacité du modèle à approcher les données d'apprentissage, et par la consistance ou capacité de généralisation à un ensemble de données distinctes des données d'apprentissage [57]. De nombreuses statistiques ont été définies dans la littérature pour évaluer ces critères [46] :

- La RMSE (*Root Mean Square Error*) qui désigne l'écart-type empirique des erreurs de prévision,
- La MAD (*Mean Absolute Deviation*) qui désigne l'écart absolu moyen entre la valeur mesurée et la valeur prédite,
- Le R^2 qui mesure la corrélation entre les valeurs mesurées et les valeurs prédites correspondantes.

Par la suite nous parlerons de RMSEC lorsque la RMSE sera estimée sur la base d'apprentissage et de RMSEP si elle est estimée sur une base de test.

Dans ce paragraphe nous avons rappelé certaines notions essentielles concernant le développement d'un modèle prédictif. Cela a pour but de faciliter la discussion sur les études qui ont été menées autour de la compréhension moléculaire et de la modélisation du PT de la coupe gazole et du VI de la coupe huile.

2.2 Compréhension moléculaire et prédiction des propriétés à froid dans les gazoles

Les propriétés à froid de la coupe gazole sont essentielles pour mesurer la qualité de ces produits. Il s'agit :

- du PT,
- de la température limite de filtrabilité (TLF),
- du PE.

2.2.1 Généralités sur les propriétés à froid de la coupe gazole

2.2.1.1 Point de trouble

Le PT (*cloud point* en anglais) est la température à laquelle les premiers cristaux apparaissent au sein de la solution lors d'un refroidissement dans des conditions normalisées. Selon Tsang *et al.* [80],

c'est le premier facteur pris en considération dans les pays froid pour la constitution des gazoles. En effet, lorsque la température est en dessous du PT, la formation de cristaux est accélérée. Les cristaux formés peuvent alors causer un dysfonctionnement des moteurs utilisant ce type de carburants. La mesure du PT est effectuée selon la norme ASTM D2500 [81] ou NF ISO EN 23015 [18]. Dans la détermination, l'échantillon est réchauffé à au moins 15°C au-dessus du PT supposé, et introduit dans un tube à essai fermé à bouchon. Un thermomètre placé à l'intérieur du tube touche le fond de celui-ci. Par la suite, on refroidit progressivement, en utilisant des bains réfrigérants de plus en plus froids, et on vérifie la limpidité du produit tous les degrés.

2.2.1.2 Température limite de filtrabilité

La TLF est la température à laquelle un volume déterminé de fluide cesse de traverser, en un temps limité, un appareil de filtration normalisé lors d'un refroidissement dans des conditions normalisées (elle est censée représenter la réalité dans un véhicule). La TLF est déterminée selon les méthodes ASTM D6371 [25] ou NF EN 116 [82]. La limite basse est de -51°C.

2.2.1.3 Point d'écoulement

Le PE d'un gazole est la température à laquelle le produit cesse de s'écouler. La mesure du point d'écoulement se fait suivant les normes ASTM D97 [27] ou NF T60-105 [21]. Le protocole de mesure est globalement le même que pour le PT excepté que le tube à essai contenant l'échantillon est analysé tous les 3°C. Cela explique la valeur particulièrement élevée (4,7°C) de l'intervalle de confiance (IC) de méthode de mesure du PE

Les fidélités liées à la détermination des propriétés à froid dans les gazoles sont précisées dans le Tableau 4.

Tableau 4 : Fidélités liées à la détermination du PT, de la TLF et du PE dans les gazoles [27]

Propriété	Répétabilité (r) NF ISO EN	Reproductibilité (R) NF ISO EN	Intervalle de confiance (IC) NF ISO EN
PT	2°C	4°C	2,8°C
TLF	1,76°C	0,102×(25-$TLF_{mesurée}$) °C	$R/\sqrt{2}$
PE	2,5 °C	6,6°C	4,7 °C

IC calculé pour un nombre d'essais égal à 1 [20]

2.2.2 Point d'écoulement des hydrocarbures purs

Pour un composé pur, le PE correspond au point de fusion. Le Tableau 5 précise les PE de différents hydrocarbures en C_{26} et montre clairement **l'influence de la structure chimique sur la cristallisation de la molécule**. En effet, pour un même nombre de carbones, on observe une variation importante du PE en fonction de la structure chimique [66] : la n-paraffine possède le PE le plus élevé ;

les iso-paraffines ont des PE inférieurs à ceux des n-paraffines. Cependant ce dernier peut varier fortement suivant la structure de l'isomère considéré. Dans le cas de l'hexacosane, le PE de ses isomères est compris entre -40 et 30°C. Les composés naphténiques et aromatiques de longueur de chaîne carbonée équivalente ont par contre des points de fusion très bas.

Tableau 5 : Variation de la tenue à froid selon la structure moléculaire des composés en C_{26} – température d'ébullition de 410°C [66]

Type de molécule	PE (°C)
n-paraffine C_{26}	56
iso-paraffine C_{26}	Entre -40 et 30
C_{26} naphténique	Environ -40
C_{26} aromatique	-60 à -30

Wise *et al.* [65] ont étudié l'évolution du PE de plusieurs paraffines en fonction de leur nombre de carbones (allant de 18 à 30). Dans le cas des n-paraffines, ils ont constaté que plus le nombre de carbones augmente, plus le PE est haut. Cette observation est aussi valable dans le cas des isoparaffines monobranchées. Ils ont noté par ailleurs que les n-paraffines ont les PE les plus élevés, suivi des isoparaffines monobranchées, puis des isoparaffines à branchements multiples.

Une étude similaire a été réalisée par Burch *et al.* [48] essentiellement sur des isoparaffines monobranchées avec un substituant méthyle, et qui ont entre 10 et 20 atomes de carbone. La variation de leur PE en fonction de la position du branchement a ainsi pu être étudiée. Ils ont observé que pour chaque type de molécule, **plus le groupement s'approche du centre de celle-ci, plus le PE diminue et moins il varie significativement,** ce qui illustre sa dépendance par rapport à la position du substituant dans le cas d'isoparaffines monobranchées.

Le cas des isoparaffines avec de multiples branchements est plus complexe. Lynch [66] a montré que le PE est également influencé par le nombre de branchements. Cette hypothèse a été nuancée par Daage [83] qui s'est appuyé sur une étude de quelques isomères du nonane et du décane**. Ce dernier souligne que le PE ne diminue pas toujours lorsque le nombre de branchements augmente**. C'est le cas par exemple du 2,7-diméthyldécane qui a un PE plus élevé (-54°C) que celui du 2-méthyldécane (-74,5°C). Il conclut de plus que **c'est plus la position du branchement qui aurait une influence prépondérante sur le PE que le nombre de branchements.**

Les études présentées ci-dessus montrent que le PE des hydrocarbures purs dépend de leur structure chimique. Les coupes gazoles étant des mélanges complexes de ces hydrocarbures, il est évident que la cristallisation des molécules présentes dans ces produits influe sur leurs propriétés à froid.

2.2.3 Compréhension moléculaire des propriétés à froid de la coupe gazole

La plupart des travaux sur la compréhension moléculaire des propriétés à froid dans les gazoles ont été axés sur la cristallisation des n-paraffines en milieu hydrocarbures. Des études ont notamment été menées au sein d'IFPEN sur la modification du processus de cristallisation des n-paraffines [84–86]. Denis et Durand [84] ont notamment identifié trois phénomènes successifs :

1. La nucléation qui se traduit par l'augmentation des forces d'attraction entre n-paraffines à basse température
2. L'accroissement en épaisseur des molécules n-paraffiniques
3. L'agglomération ou migration des molécules cristallisées pour la formation de cristaux visibles.

La structure symétrique des n-paraffines est selon Turner [87] un des facteurs principaux permettant d'expliquer leurs températures de fusion particulièrement hautes. Ce dernier a également souligné que la cristallisation des molécules d'hydrocarbures dépend du milieu moléculaire. Ces propos ont été appuyés par différents travaux [67,88]. A partir d'une étude de mélanges de distillats moyens (allant d'essences lourdes à gazoles lourds), Rossemyr [88] a notamment montré qu'il existe une forte corrélation positive entre la proportion de n-paraffines dans les mélanges et le PT, et que cette corrélation varie suivant les proportions des différentes coupes. Ces observations sont également valables pour la TLF [88].

Krishna *et al.* [67] ont réalisé le même type d'étude sur le PE de différentes coupes gazoles étroites. Les graphes obtenus sont représentés sur la Figure 4. On note que plus **le point de coupe final est élevé, plus la valeur du PE est grande**. Les auteurs ont également souligné que **les n-paraffines de longueur de chaîne inférieure à 15 ont une influence moindre sur le PE** des coupes analysées.

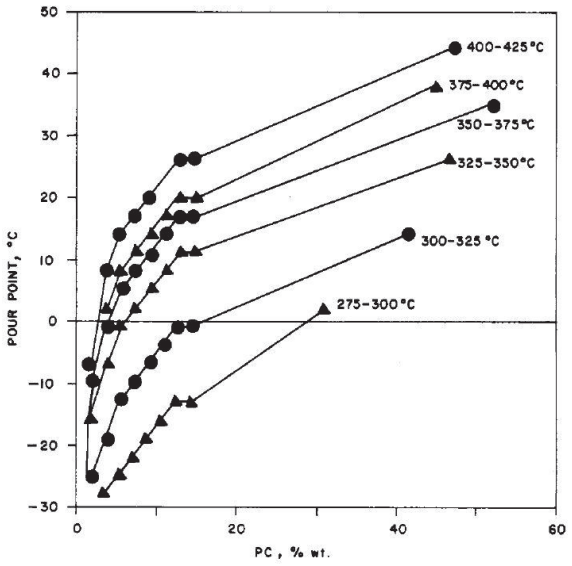


Figure 4 : Effet de la concentration en n-paraffines (PC) sur le point d'écoulement de coupes étroites de gazoles [67]

2.2.4 Modélisation des propriétés à froid dans les gazoles

De nombreuses études ont été effectuées sur la modélisation des propriétés à froid des gazoles. Outre les approches basées sur des équations thermodynamiques, des modèles de corrélation faisant intervenir des caractéristiques moléculaires diverses ont été proposés. Ces caractéristiques des coupes gazoles ont été obtenues à partir de données analytiques issues de différentes techniques :

- des techniques spectroscopiques telles que la Résonance Magnétique Nucléaire (RMN) du ^{1}H ou du ^{13}C [34], l'infrarouge (IR) [30] et la spectrométrie de masse (MS) [89] ;
- des techniques chromatographiques telles que la chromatographie gazeuse monodimensionnelle (GC) et bidimensionnelle (GC×GC) [35] et la chromatographie en phase liquide haute performance (HPLC) [38] ;
- des méthodes de mesure standard de propriétés globales (densité, indice de cétane, *etc.*) [3,90].

2.2.4.1 Approche thermodynamique pour la prédiction des propriétés à froid dans les gazoles

Des modèles thermodynamiques ont été proposés [80,91]. Reddy [91] a développé un modèle de prédiction du point de trouble (PT). La méthode consiste à assimiler la coupe gazole à un solvant dans lequel est placée une n-paraffine de référence. Celle-ci est choisie en fonction des fractions molaires des différentes n-paraffines présentes dans le mélange. Le PT est alors déterminé par application de l'équation dite de « solubilité » à cette n-paraffine de référence :

$$\ln N = -\frac{\Delta H_m}{R}\left(\frac{1}{T} - \frac{1}{T_m}\right) \qquad \textbf{(Eq.1. 2)}$$

où N, ΔH_m et T_m représentent respectivement la fraction molaire, l'enthalpie de fusion et la température de fusion d'un composé, R est la constante des gaz parfaits. Sur 8 échantillons analysés pour un PT variant entre -12°C et 11°C, une erreur absolue moyenne de 1°C a été obtenue par rapport à la méthode de référence NF EN ISO 23015 [18]. Une approche similaire a été proposée par Tsang *et al.* [80] pour la prédiction du PT de mélanges de distillats allant d'essences légères à des gazoles lourds. Ces différents modèles obtenus à partir de mélanges dont les PT sont globalement compris entre -38°C et -1°C ont fourni une erreur absolue moyenne autour de 0,40°C.

Ces études semblent montrer que les modèles thermodynamiques sont performants. Toutefois, la signification statistique de ces approches peut être questionnée car très peu d'échantillons ont été analysés et les modèles n'ont été évalués que sur les données d'apprentissage, **ce qui ne permet pas de garantir leur consistance.** Par ailleurs, **les enthalpies de fusion des n-paraffines et les proportions des mélanges tels qu'ils sont définis dans l'étude de** Tsang *et al.* [80] **ne sont pas toujours accessibles en raffinerie.**

2.2.4.2 Modèles de régression linéaire (généralisée) ou modèles corrélatifs pour la prédiction des propriétés à froid dans les gazoles

Des modèles corrélatifs pour la prédiction des propriétés à froid des gazoles faisant intervenir d'autres propriétés de produits ont été développés. Deux modèles de prédiction du PT et du PE des

gazoles ont été publiés entre 1988 et 1990 [92–94]. Un premier modèle qui exprime la corrélation entre les propriétés à froid et la concentration en n-paraffines. Un second modèle tient en plus compte des concentrations en hydrocarbures aliphatiques autres que les n-paraffines et en aromatiques. Des erreurs quadratiques moyennes d'environ 3,4°C ont été obtenues à chaque fois pour 40 échantillons. Ces modèles ont été développés pour des coupes gazoles standard. Par la suite, Cookson *et al* [95] ont proposé des modèles plus généraux pour des coupes gazoles à intervalle d'ébullition variable (avec un point d'ébullition initial compris entre 230°C et 250°C et un point d'ébullition final situé entre 320°C et 370°C), mettant en avant la nécessité d'introduire une information relative aux points de coupe [95]. Les modèles proposés pour le PT et le PE sont donnés dans les équations suivantes :

$$\boldsymbol{PT = a_1 C_{np} + a_2 C_a + a_3 T_{10} + a_4 T_{90} + a_5} \qquad \textbf{(Eq.1. 3)}$$

$$\boldsymbol{PE = b_1 C_n - b_2 C_a + b_3 T_{10} + b_4 T_{90}} \qquad \textbf{(Eq.1. 4)}$$

où les a_i et les b_i sont les paramètres du modèle, et C_{np} et C_a désignent respectivement les pourcentages de carbone n-paraffiniques et aromatiques estimés à partir de données RMN du ^{13}C. Les T_i sont les températures de la distillation simulée (DS) des gazoles analysés faite par chromatographie en phase gazeuse. Il s'agit de la température à laquelle i% de la masse du produit analysé s'est évaporée lors d'une augmentation régulière de la température (paragraphe 3.1.2).

Au sein d'IFPEN, Dulot *et al.* [12] ont proposé les modèles suivants :

$$\boldsymbol{PT = \alpha_1 T_{10} + \alpha_2 C_p^2 + \alpha_3 {T_{90}}^2 + \alpha_3} \qquad \textbf{(Eq.1. 5)}$$

$$\boldsymbol{TLF = \beta_1 PT + \beta_2} \qquad \textbf{(Eq.1. 6)}$$

$$\boldsymbol{PE = \gamma_1 PT + \gamma_2} \qquad \textbf{(Eq.1. 7)}$$

où les α_i, β_i et γ_i désigne leurs paramètres respectifs. Le pourcentage de carbone paraffinique (C_p) est déterminé par méthode dite n-d-M [22]. Le PE et la TLF sont déterminés à partir du PT (ces propriétés étant très corrélées entre elles. Les performances de ces modèles qui ont été évaluées sur 185 et 198 gazoles sont données dans le Tableau 6. Des erreurs absolues moyennes proches ou en dessous des incertitudes liées aux mesures de référence ont été obtenues sur la base d'apprentissage. **Si la précision de ces modèles est plus ou moins satisfaisante, leur consistance liée en particulier à la largeur des coupes n'a pas été démontrée.**

Tableau 6 : Statistiques des modèles de prédiction proposés par Dulot *et al.* [12]

Corrélation	Nombre d'échantillons	Min (°C)	Max (°C)	MAD	Incertitude méthode NF ISO EN
PT	198	-27	2	3,34°C	2,8°C
PE	189	-45	3	3,07°C	4,7°C
TLF	185	-37	11	1,45°C	0,07×(25-$TLF_{mesurée}$) °C

Plus récemment, des travaux de prédiction des propriétés à froid des gazoles à partir de la chromatographie en phase gazeuse bidimensionnelle (GC×GC) ont été effectués par Souchon *et al.* [96]. Un des avantages majeurs de cette technique de caractérisation est qu'elle permet de distinguer les n-paraffines des isoparaffines malgré leurs structures très voisines. Pour cette étude, une base d'échantillons constituée de 48 gazoles a été collectée, puis analysée. A partir de ces données, ils ont estimé un certain nombre de descripteurs potentiels parmi lesquels :

- La T_{90} et la T_{95} obtenues par DS
- Les teneurs par famille d'hydrocarbures identifiés
- Le Ratio n/iso paraffines
- Le nombre moyen de carbones pour les n-paraffines
- L'indice de paraffine (PI) défini par :

$$\boldsymbol{PI = \sum_{n-paraffines} n \times x_n} \qquad \textbf{(Eq.1. 8)}$$

où n est le nombre de carbone et x_n la fraction massique de la n-paraffine. Plusieurs modèles (linéaires et quadratiques) développés à partir de ces propriétés ont été étudiés, mais les résultats montrent qu'ils sont peu performants et manquent de consistance [96].

L'avantage des modèles basés sur des corrélations entre propriétés est leur interprétabilité. En effet, dans le domaine pétrolier, l'objectif de la modélisation n'est pas seulement de prévoir. Modéliser peut également permettre de mieux comprendre les phénomènes qui ont lieu durant les procédés ou de confirmer des tendances empiriques. Par exemple, si l'on analyse les coefficients des différents modèles présentés dans ce paragraphe, on peut notamment tirer les observations suivantes : (1) les propriétés à froid des gazoles augmentent avec le pourcentage de carbone paraffinique et les points de coupe ; (2) la présence de carbone aromatiques tend à diminuer les propriétés à froid.

Trouver ces différentes corrélations n'est cependant pas toujours simple, notamment dans le cas de modèles polynomiaux comme ceux qui ont été développés par Dulot *et al.* [97], et l'augmentation du nombre de descripteurs expose au risque d'informations redondantes.

2.2.4.3 Modèles chimiométriques de prédiction des propriétés à froid dans les gazoles

Des travaux de modélisation des propriétés à froid de la coupe gazole ont été effectués au sein d'IFPEN, notamment par Quignard *et al.* [69,70]. Des gazoles provenant de différents procédés (HDT, HCK, *Fluid Catalytic Cracking, Mild Hydrocracking*, *etc.*) ont été analysés pour la prédiction du PT, de la TLF et du PE. Différents types de modèle (PLS et PCR) ont été développés à partir de données issues de la DS [98], (T_i, $i = 5, 10, \dots, 95$) et / ou de la composition chimique des coupes obtenue par MS [99]. Selon les auteurs, les performances de ces modèles sont globalement éloignées des précisions des mesures de référence.

Au cours d'une même étude présentée au paragraphe précédent, Souchon *et al.* [96] ont développé un modèle des modèles de prédiction des propriétés à froid des gazoles par régression PLS

appliquée aux données GC×GC. Les résultats obtenus sont récapitulés dans le Tableau 7. Là encore, les valeurs de RMSEC et de RMSEP étant relativement élevées.

Tableau 7 : Statistiques de performances des modèles développés par Souchon *et al.* [100]

Propriété modélisée	Min / Max Apprentissage (°C)	Min / Max Test (°C)	RMSEC (°C)	RMSEP (°C)
PE	-43 / 17	-38 / -10	5,5	5,8
TLF	-40 / 5	-40 / 5	5,5	5,4
PT	-36 / 8	-32 / 0	5,0	4,5

2.2.4.4 Modèles de prédiction des propriétés à froid dans les gazoles basés sur les réseaux de neurones

La complexité de la modélisation des propriétés à froid des gazoles a conduit certains auteurs [101–103] à développer des modèles de prédiction basés sur la méthode des réseaux de neurones. Marinovic *et al.* [101] ont proposé des modèles de prédiction pour le PT et la TLF prenant en entrée l'indice de cétane, la viscosité, les données de la DS, *etc.* Sur 180 échantillons analysés, des MAD de 0,58°C et 1,46°C ont été obtenues respectivement pour chaque propriété. Toutefois, ces modèles n'ont pas été évalués sur une base de données indépendante. Dans le cas des réseaux de neurones le risque de surapprentissage est d'autant plus élevé qu'on dispose d'une base d'échantillons relativement limitée. Cela augmente la nécessité d'effectuer une étape de test (paragraphe 2.1.1).

L'étude proposée par Pasadakis *et al.* [102] est plus intéressante de ce point de vue. Les modèles développés pour la prédiction du PT et du PE ont pour entrées des composantes principales formées à partir de spectres IR pour une région spectrale comprise entre 1700 cm^{-1} et 600 cm^{-1}. Dans cette région on peut observer les bandes d'absorption suivantes [30] :

- élongations des doubles liaisons C=C (1650 – 1550 cm^{-1}) ;
- déformations des liaisons C-H (1500 – 1300 cm^{-1}) ;
- déformations des liaisons C-H hors du plan (950 – 700 cm^{-1}) qui peuvent être attribuées à des cycles aromatiques.

La base de données a été divisée en deux : une base d'apprentissage (85% des échantillons) et une base de test. Une erreur absolue moyenne de prédiction de 2,5°C a été obtenue pour le PT et de 2,3°C pour le PE. Si ces résultats sont intéressants du point de vue de la précision et de la consistance, ce modèle pose toutefois deux difficultés d'un point de vue des objectifs de cette thèse : la première est que la composition de réseaux de neurones et de la régression PLS débouche sur un modèle difficilement compréhensible ; la seconde est que pour les modèles du simulateur IFPEN, l'utilisation de données Infrarouge est inenvisageable.

Des travaux de compréhension moléculaire et de prédiction du VI de la coupe huile ont également été recensés. Ils sont présentés ci-dessous.

2.3 Compréhension moléculaire et prédiction du VI dans les huiles de base

Ce paragraphe présente les différents travaux qui ont été effectués sur la compréhension moléculaire et la modélisation du VI de la coupe huile. Les généralités sur cette propriété sont préalablement rappelées.

2.3.1 Généralités sur le VI des huiles de base

Le VI caractérise la variation de la viscosité d'une huile avec la température [3]. Il est calculé à partir de la mesure de la viscosité du fluide à deux températures de référence (généralement à 40 et 100°C). Le VI est issu d'une échelle logarithmique caractérisée par deux huiles de référence : une huile très naphténique (L) issue du golfe du Mexique qui montrait une grande variation de la viscosité avec la température (le VI de référence a été fixé à 0 pour cette huile) ; une huile paraffinique (H) qui présentait une faible variation par rapport à la température (le VI de référence a été fixé à 100 pour cette huile). Ainsi, les échantillons qui possèdent une variation de viscosité avec la température inférieure à celle de H auront donc un VI supérieur à 100 et inversement, une huile possédant une variation de viscosité par rapport à la température supérieure à celle de L, aura un VI négatif. Le calcul du VI est défini par la norme européenne NF ISO 2909 [19] qui est équivalente à la norme ASTM D2270 [104]. La procédure de calcul de l'indice de viscosité est la suivante [19] :

- Si U>H (*i.e.* VI<100)
    - $VI = \frac{L-H}{L-U} \times 100$
- Si U<H (*i.e.* VI>100)
    - $VI = \frac{(10^N - 1)}{0{,}00715} + 100$ avec $N = \frac{\log H - \log U}{\log Y}$

Où :

- U est la viscosité cinématique (mm²/s) à 40°C d'un produit pétrolier dont le VI est à mesurer ;
- L est la viscosité cinématique (mm²/s) à 40°C d'un produit pétrolier dont l'indice de viscosité est 0 et ayant la même viscosité cinématique à 100°C que le produit pétrolier dont l'indice de viscosité est à mesurer ;
- H est la viscosité cinématique (mm²/s) à 40°C d'un produit pétrolier dont l'indice de viscosité est 100 et ayant la même viscosité cinématique à 100°C que le produit pétrolier dont l'indice de viscosité est à mesurer ;
- Y est la viscosité cinématique (mm²/s) à 100°C du produit pétrolier dont l'indice de viscosité doit être déterminé.

Les valeurs de L et H sont données dans les normes pour des viscosités cinématiques à 100°C comprises entre 2 et 70 mm²/s. Un exemple de procédure de calcul du VI est donné en Annexe C.
Les valeurs de fidélités sont données pour la norme dans le Tableau A. 6 pour la norme NF ISO 2909 [19]. Pour cette dernière, l'IC dépend des valeurs de VI et de viscosité à 100°C. Pour la norme ASTM D2270 [104], le VI est mesuré avec une reproductibilité de 2 points.

2.3.2 VI et point d'écoulement des huiles de base

Pour les huiles de base, le point d'écoulement (PE) est généralement compris entre -9 et -24°C. L'obtention des PE bas est nécessaire pour assurer la lubrification par temps froid mais également pour des applications de type réfrigération [3]. La norme ASTM D4304 [24], relative aux spécifications des huiles minérales et synthétiques pour les moteurs requiert un PE inférieur à -6°C. La mesure du PE se fait suivant les normes ASTM D97 [27] ou NF T60-105 [21]. Pour les huiles de base, cette propriété est très importante, notamment pour lors de la mesure du VI. En effet, ce dernier doit être déterminé à iso PE. Au sein d'IFPEN, le PE est mesuré sur l'huile après déparaffinage au solvant à -20°C. L'intervalle de confiance de la méthode est de ±5,7°C. Le PE est donc considéré aux spécifications lorsque sa valeur est comprise entre -14 et -26°C.

2.3.3 VI des hydrocarbures purs

La détermination du VI de nombreux hydrocarbures purs de différentes familles a précédemment été reportée au sein du projet API 42 en 1967 [105]. Les résultats de cette étude ont notamment été exploités par de nombreux auteurs [66,83,106] et sont en partie illustrés sur la Figure 5. Ils montrent que les n-paraffines ont les VI les plus élevés (le VI moyen des n-paraffines présents sur la Figure 5 est de 175). Ils soulignent par ailleurs que les isoparaffines monobranchées et les mononaphtènes avec de longues chaînes alkyles ont également des hauts VI (les VI moyens de ces composés sont respectivement de 155 et 142). Enfin, les VI les plus bas correspondent en général aux dinaphtènes et aux composés aromatiques.
La chute de VI qui est observé au-delà de n-paraffines C_{35} est sans doute due au fait que ces n-paraffines ne sont pas liquides en dessous de 100°C [66]**. Par ailleurs, les n-paraffines ont des points d'écoulement très élevés et ne peuvent donc pas être conservées dans la fraction huile pour respecter les spécifications**. Cela explique la nécessité d'effectuer un déparaffinage de la fraction non convertie en sortie du réacteur HCK (Figure 3) pour obtenir une huile qui réponde aux spécifications. Les isoparaffines et les mononaphtènes ont au contraire des PE relativement bas. La Figure 5 permet également de noter que pour une famille d'hydrocarbures donnée, le VI peut varier significativement suivant la structure de la molécule. On note d'une part que le VI augmente globalement avec la longueur de la chaîne carbonée et d'autre part qu'il diminue avec le nombre de cycles pour les composés naphténiques et aromatiques. Ces observations ont été confirmées par Lynch [66] puis par Daage [83]. Ces derniers ont également montré que la structure des isomères (nature et position du substituant et nombre de branchements) a une influence sur le VI de ce type des hydrocarbures. Une étude de différents

isomères du n-tétracosane (n-C_{24}) a également été réalisée. A la suite de cette étude, Lynch [66] a émis trois observations : (1) le VI diminue lorsque le nombre de branchement augmente ; (2) le VI est plus élevé lorsque le substituant est une chaîne alkyle que lorsqu'il s'agit d'un cycle naphténique, les VI les plus faibles étant pour des substituants phényles ; (3) pour un substituant donné, le VI diminue lorsque la position du branchement se rapproche du centre de la molécule.

Au cours d'une dernière étude réalisée cette fois sur plusieurs hydrocarbures de différentes familles (alkyl alcanes, cyclopentyl alcanes, cyclohexyl alcanes et phenyl alcanes), Lynch [66] a souligné que les hydrocarbures qui ont à la fois des VI supérieurs à 100 et des PE en dessous de 0°C (caractéristiques d'une huile de bonne qualité) sont **majoritairement des composés possédant de longues chaînes avec des ramifications courtes contenant un cycle naphténique à cinq atomes de carbone plutôt qu'un cycle aromatique**.

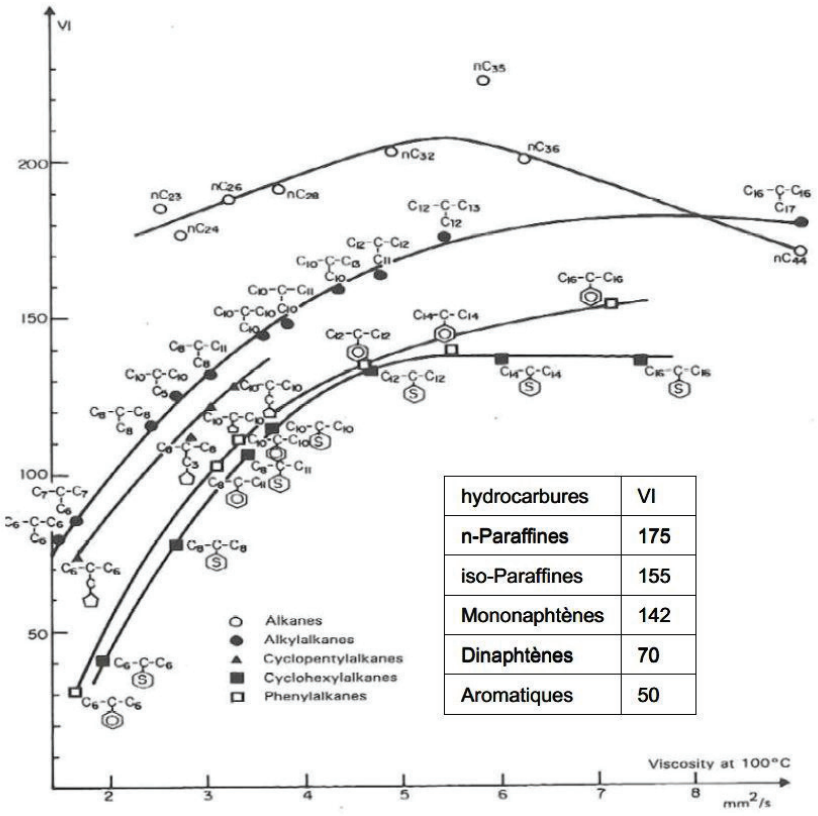


Figure 5 : VI de différents hydrocarbures purs en fonction de leur viscosité à 100°C [66]

En résumé, le VI est fortement influencé par la nature chimique des molécules présentes dans les fractions huiles. Un classement des types de molécule suivant l'ordre décroissant de leur VI serait le suivant :

1. n-paraffines,
2. isoparaffines monobranchées,

3. isoparaffines multibranchées et mononaphtènes avec de longues chaînes alkyles
4. monoaromatiques avec de longues chaînes alkyles
5. polynaphtènes, polynaphténoaromatiques, polyaromatiques

La longueur de la chaîne (plus le nombre de carbones augmente plus le VI augmente), le taux de branchements de la molécule (plus il est faible, plus le VI est grand) et la position des substituants (une ramification en début ou en fin de chaîne donne un meilleur VI que dans le cas d'une ramification en milieu de chaîne), influencent le VI. Pour une molécule d'hydrocarbure, le nombre de paramètres qui influent sur le VI est important. De plus, les huiles de base sont des mélanges complexes d'hydrocarbures purs. Des techniques analytiques poussées ont donc été utilisées pour la caractérisation des huiles de base.

2.3.4 Compréhension moléculaire du VI dans les huiles de base

La plupart des études sur la compréhension moléculaire du VI des huiles de base font appel à des analyses spectroscopiques, principalement la RMN du ^{13}C [73,107–109] et la spectroscopie IR [71,110–112].

2.3.4.1 Compréhension moléculaire du VI par RMN

La spectroscopie RMN est une technique qui exploite les propriétés magnétiques de certains noyaux atomiques [3]. Elle est basée sur le phénomène de résonance magnétique nucléaire. Les applications les plus courantes pour l'étude de composés organiques sont la RMN du ^{1}H et du ^{13}C [34]. Les pics observés sur les spectres sont caractéristiques de l'environnement à l'échelle atomique du noyau observé. Cette technique est donc particulièrement adaptée à la caractérisation de la structure moléculaire des composés.

En 1996, Sarpal *et al.* [113] ont réalisé une étude de caractérisation des hydrocarbures présents dans les huiles de base, basée à la fois sur la RMN du ^{13}C et sur la RMN 2D (couplage entre RMN du ^{1}H et RMN du ^{13}C). L'exploitation des spectres a permis d'identifier différentes structures moléculaires relatives aux espèces présentes dans les huiles. Ces identifications ont été plus récemment complétées par Sperber *et al.* [114] *via* l'analyse RMN du ^{13}C de cires paraffiniques. Des exemples de structures moléculaires identifiables par spectroscopie RMN du ^{13}C sont donnés en Annexe E. Ces auteurs ont également proposé des formules pour l'estimation de paramètres structuraux moyens à partir des intensités des pics, notamment :

- les proportions en n-paraffines et en iso-paraffines,
- la proportion de carbone n-paraffiniques, isoparaffiniques, naphténiques et aromatiques,
- la longueur moyenne des chaînes alkyles,
- le nombre de sites de branchements moyens.

Ces travaux ont servi de base pour de nombreuses études. Sarpal *et al.* [113] ont comparé les spectres RMN du ^{13}C d'huiles obtenues par différents procédés : des huiles obtenues par hydrofinition qui ont un VI compris entre 90 et 110 et des huiles obtenues par HCK qui ont un VI compris entre 120 et 150.

Ils ont noté que les spectres des huiles issues d'HCK présentent un signal intense dans les zones correspondant à des structures isoparaffiniques monobranchées et un signal faible dans celles qui correspondent à des structures qui présentent des branchements multiples.

Au cours de cette même étude, la corrélation entre le VI et des paramètres structuraux moyens a été étudiée. Ces travaux ont notamment été exploités par Verdier *et al.* [73] qui ont représenté la variation du VI en fonction de ces paramètres pour 37 huiles analysées sur la Figure 6. Ils ont observé que : la proportion en carbones paraffiniques (C_p), les proportions en n et iso paraffines et la proportion molaire de branchement méthyl ont globalement tendance à augmenter le VI (Figure 6a, b, e et f). Ils ont également noté que la proportion en carbone naphténique (C_n) tend à diminuer le VI (Figure 6c). Ces résultats sont en adéquation avec les conclusions faites sur le VI des hydrocarbures purs. Cependant, un point d'interrogation subsiste concernant les branchements. En effet, on note que le VI diminue lorsque le nombre de sites de branchement augmente (Figure 6d) tandis que le pourcentage molaire de branchements méthyles a un effet positif sur cette propriété (Figure 6f). **De plus certaines des tendances observées semblent clairement partielles et refléter en premier lieu les différences de composition relatives aux procédés d'obtention des huiles analysées**.

Verdier *et al.* [73] ont par ailleurs élargi leur étude à des DSV et à des effluents issus d'HDT et d'HCK. A partir des 20 échantillons analysés par RMN du ^{13}C, ils ont identifié 3 pics qui ont une influence positive sur le VI (Figure 7) : le pic à 27,3 ppm qui correspond à un méthylène (CH_2) en position β par rapport à un branchement méthyle ; les pics à 29,9 et 32,0 ppm qui correspondent respectivement à des CH_2 en position δ ou plus et γ d'une chaîne droite (structure EEEn et S_3, Annexe E). Ils ont également souligné que les pics à 11,4 ppm qui correspondent à un carbone primaire (CH_3) à la fin d'un branchement éthyle (structure $1B_2$, Annexe E) et ceux qui sont compris entre 117 et 150 ppm qui correspondent à des carbones aromatiques ont une influence négative sur le VI. Ces observations confirment que la longueur de chaîne carbonée et la faible présence de branchements tendent à augmenter le VI. Elles montrent également l'impact négatif des aromatiques sur cette propriété.

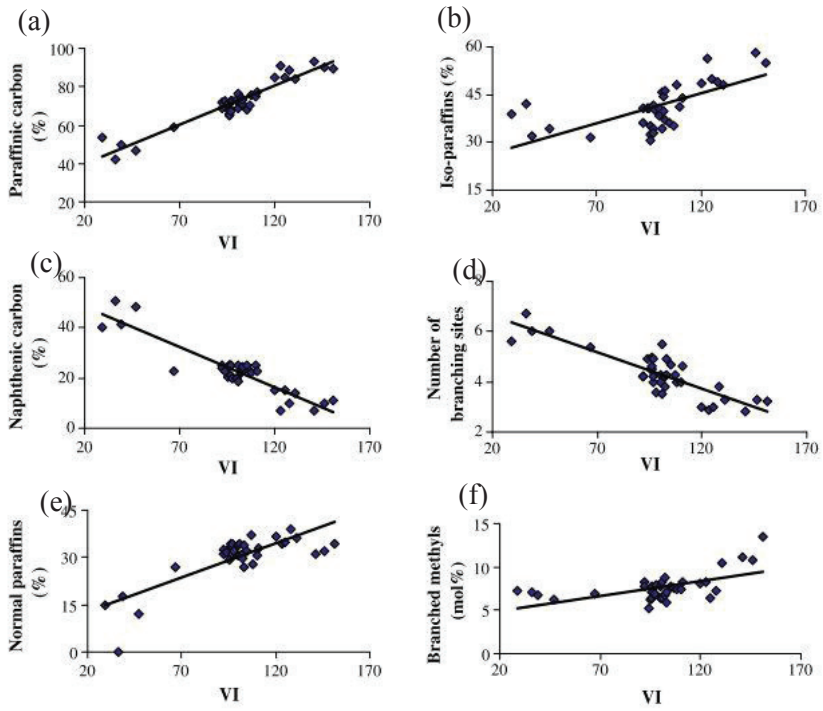


Figure 6 : Evolution du VI en fonction a) du pourcentage de carbones paraffiniques ; b) de la proportion en isoparaffines ; c) du pourcentage de carbones naphténiques, d) du nombre de sites de branchement ; e) de la proportion en n-paraffines ; f) de la proportion molaire de branchement de type méthyle [73]

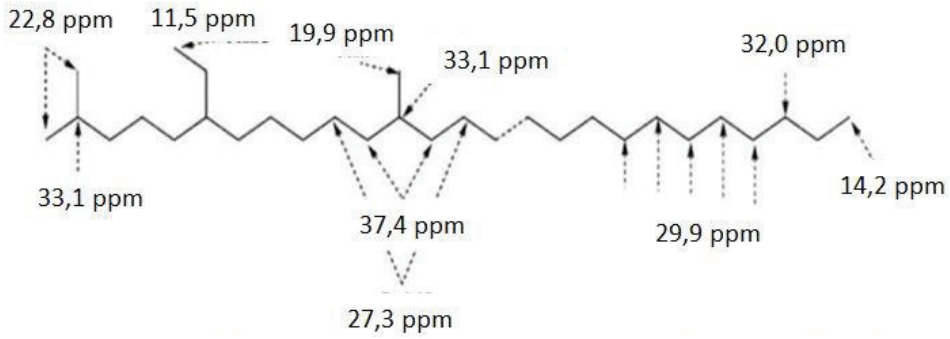


Figure 7 : Correspondance entre certains déplacements chimiques et les structures moléculaires correspondantes [73]

2.3.4.2 Compréhension moléculaire du VI par spectroscopie Infrarouge

La spectroscopie IR est une technique basée sur l'analyse de la région infrarouge du spectre électromagnétique (12500 à 400 cm^{-1}). Dans cette région les bandes d'absorption observées sont principalement dues aux vibrations des liaisons des molécules (4000 à 400 cm^{-1}). Généralement les bandes observées sont spécifiques et peuvent être pour la plupart attribuées à un groupement chimique. La spectroscopie IR est donc particulièrement bien adaptée à l'identification de composés organiques ou de différents types de liaison dans les molécules.

Sastry *et al.* [71] ont effectué une étude sur des huiles produites à partir de différents procédés. Les spectres infrarouges de trois d'entre elles sont représentés sur la Figure 8. Ils ont observé que la bande à 1600 cm^{-1} (caractéristique d'une liaison C=C) n'est pas présente sur le spectre de l'huile issue de l'HCK (Figure 8a). Cela s'explique par la quasi-absence d'aromatiques dans cette huile. Les bandes caractéristiques sont différentes en fonction du VI de l'échantillon. En effet, les bandes caractéristiques des aromatiques observées à 1600 cm^{-1} et 815 cm^{-1} (déformation des liaisons CH hors du plan) entourées respectivement en rouge et vert sur la Figure 8 sont uniquement présentes sur les spectres des huiles possédant des VI de 47 et 100 (Figure 8a et c). *A contrario*, la bande à 720 cm^{-1} (zone entourée en bleu) qui caractérise des paraffines n'est détectée que dans les huiles à VI élevé [115]. Ces observations confirment l'impact positif de l'absence de composés aromatiques et de la présence de paraffines sur le VI.

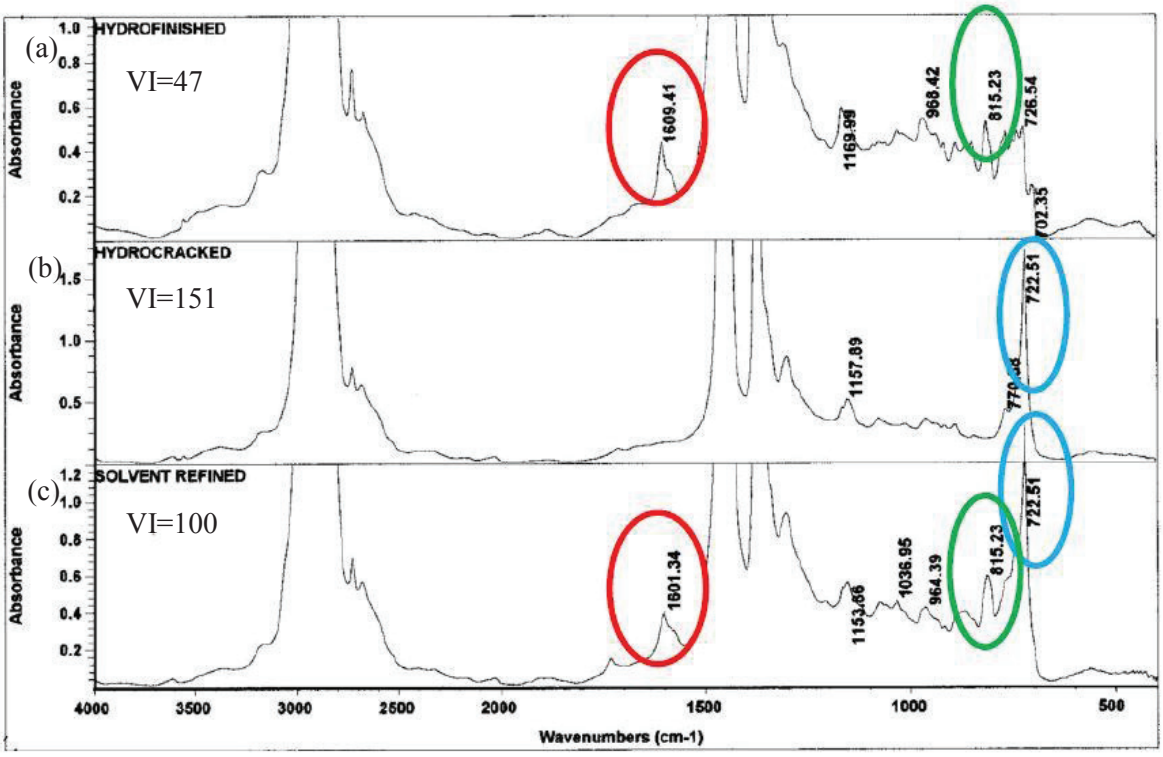


Figure 8 : Spectres ATR-IR d'huiles de base produites par différents procédés a) huile obtenue par hydrofinition ; b) huile obtenue par hydrocraquage ; c) huile obtenue par raffinage au solvant [71]

Des travaux ont également été effectués à IFPEN autour de la faisabilité de l'étude des variations du VI par IR [14]. **Cette étude basée sur la caractérisation de 8 coupes UCO issues d'HCK et de VI allant de 100 à 140, a montré qu'il était difficile d'observer des différences significatives entre les spectres obtenus.**

2.3.5 Modélisation du VI dans les huiles de base

Comme dans le cas des propriétés à froid des gazoles (paragraphe 2.2.4), l'exploitation de données issues de techniques analytiques telles que la RMN et l'IR a permis le développement de modèle de prédiction du VI.

2.3.5.1 Modèles de régression linéaire (généralisé) pour la prédiction du VI dans les huiles

Kobayashi *et al.* [116] ont proposé un modèle de prédiction du VI d'huiles de base produites par HCK de cires Fischer-Tropsch (charges particulièrement riches en paraffines). Le modèle proposé est le suivant :

$$\boldsymbol{VI = 0,0008x^2 + 0,7559x - 17,449} \qquad \textbf{(Eq.1. 9)}$$

avec :

$$\boldsymbol{x = \frac{ACN^2}{ABN}} \qquad \textbf{(Eq.1. 10)}$$

ACN (*Average Carbon Number*) et ABN (*Average Branching Number*) désignant respectivement le nombre d'atomes de carbone moyen et le nombre de branchement moyen des molécules présentes dans les 8 huiles analysées. Ces propriétés structurales ont été estimées à partir des spectres RMN du ^{13}C de ces huiles. Un coefficient de corrélation de 0,99 a été obtenu pour ce modèle. Bien que ce résultat soit intéressant, il est difficilement généralisable. En effet, la spécificité des Cires Fischer-Tropsch simplifie l'estimation des grandeurs (ACN et ABN) mises en jeu.

Sharma *et al.* [117,118] ont publié deux études de prédiction du VI d'huiles de groupe II et III (99 < VI < 132). Là encore les données d'entrées sont des paramètres structuraux estimées à partir des spectres RMN du ^{13}C de ces huiles. Le modèle suivant a notamment été proposé :

$$\boldsymbol{VI = 2480 - 22,7C_{np} - 23,9C_{ip} - 37,0C_{N,CH_3} - 22,5C_{N,others}} \qquad \textbf{(Eq.1. 11)}$$

Les entrées sont les pourcentages de carbones :

- n-paraffiniques (C_{np}),
- isoparaffiniques (C_{ip}),
- naphténiques primaires (groupe méthyle substitué à un carbone naphténique, noté C_{N,CH_3})
- naphténiques de tout autre type ($C_{N,others}$).

Ce résultat obtenu sur un nombre très limité d'échantillon (6) a été remis en question plus récemment par Sarpal *et al.* [109] à partir d'une étude d'huiles de groupe I à III (90 < VI < 150). Ils ont par ailleurs proposé un modèle alternatif basé sur l'intégration de pics caractéristiques du spectre RMN du ^{13}C (ces pics font référence à des structures bien identifiées). Le modèle suivant a ainsi été proposé :

$$\boldsymbol{VI = 65,76 + 4,15S_2 + 1,62S_{6+7}} \qquad \textbf{(Eq.1. 12)}$$

où S_2 désigne l'aire (normalisée) du pic à 22,8 ppm qui fait référence à un groupe méthyle branché sur un carbone en position β par rapport à un carbone en bout de chaîne, et S_{6+7} désigne l'aire du pic à 27,0 ppm caractéristique d'un branchement méthyle sur un carbone en milieu de chaîne [72]. Une RMSEC

de 2,49 a été obtenue pour 13 huiles analysées. Là encore, malgré une précision intéressante, le nombre réduit d'échantillon du modèle sur une plus large gamme d'huile est à vérifier.
Les travaux de Verdier *et al*. [73] ont abouti à un modèle similaire :

$$\boldsymbol{I = 203 - 41,32\left(\frac{A_2}{A_3} + \frac{A_7}{A_3} + \frac{A_{10}}{A_3}\right) - 2,206\ 10^{-3}\left(\frac{A_{aro}}{A_3}\right)^5} \qquad \textbf{(Eq.1. 13)}$$

où A_2, A_3, A_7 et A_{10} sont les aires respectives des pics à 11,5 ppm, 14,1 ppm, 19,9 ppm et 22,8 ppm (voir structures correspondantes sur la Figure 7) et A_{aro} désigne l'aire totale des pics répertoriés dans la zone correspondant aux carbones aromatiques (entre 117 et 149 ppm). Sur 16 échantillons analysés (de VI compris entre -104 et 146), une MAD de 3,0 a été obtenue. Suite à ces travaux, une étude de prédiction du VI de coupes UCO et huile a d'ailleurs été réalisée au sein d'IFPEN. Un modèle analogue à celui proposé par Verdier *et al.* [73] a ainsi été développé à partir de 25 échantillons de VI compris entre 60 et 140. Sur ces données d'apprentissage, une RMSEC de 4,3 a été obtenue sur les données d'apprentissage. De manière générale, les statistiques moyennes du modèle proposé par Verdier *et al.* [73] peuvent traduire une difficulté du modèle à s'adapter à des échantillons d'une telle diversité (DSV, DSV hydrotraités, DSV hydrocraqués, UCO, huiles). Par contre, cette diversité d'échantillons analysés et la large gamme de VI couverte donnent un intérêt particulier à ces études.

Toujours au sein d'IFPEN, des études sur la prédiction du VI avaient été effectuées d'abord par Billon *et al.* [119], puis par Dulot [12]. Elles ont abouti au développement au modèle suivant :

$$\boldsymbol{VI_{out} = VI_{in} + K \times X_{370+}(X_{370+} - (a \times VI_{in} + b))} \qquad \textbf{(Eq.1. 14)}$$

où VI_{in} et VI_{out} sont respectivement les indices de viscosité des coupes déparaffinées en entrée et en sortie de l'étape considérée, X_{370^+} désigne le taux de conversion nette par passe en UCO 370^+ (Figure 3) et K, a et b sont les paramètres à estimer. X_{370+} mesure la variation relative de la part de coupe pétrolière dont l'intervalle d'ébullition se situe au-dessus de 370°C entre la charge (notée $C_{370^+,charge}$) et liqtot ($C_{370^+,liqtot}$). Elle est donnée par la formule suivante [7] :

$$\boldsymbol{X_{370^+} = \frac{C_{370^+,charge} - C_{370^+,liqtot}}{C_{370^+,charge}}} \qquad \textbf{(Eq.1. 15)}$$

Dans le cas des huiles produites par HDT, on note que la MAD et la RMSEC sont respectivement de 3,4 de 4,6. Pour les huiles issues de l'HCK, ces valeurs sont respectivement de 4,6 et de 5,6. Ces statistiques qui ont été évaluées sur les données d'apprentissage sont encore insuffisantes et soulignent l'intérêt de nos travaux.

2.3.5.2 Modèles chimiométriques pour la prédiction du VI dans les huiles

Sastry *et al.* [71] ont développé un modèle PLS de prédiction du VI à partir de données spectrales obtenues par analyses ATR-IR (*Attenuated Total Reflectance InfraRed*). Deux zones spectrales ont été retenues pour ce modèle :

1. la zone comprise entre 3100 et 2700 cm^{-1} (élongations des liaisons C-H dans les groupements CH_2 et CH_3 et des liaisons =C-H) ;

2. la zone comprise entre 1800 et 650 cm^{-1} (élongations des liaisons C=C, déformations des liaisons C-H dans et hors du plan).

La base de données constituée de 60 échantillons (29 < VI< 151) a été divisée en une base d'apprentissage (35 échantillons) et une base de test (25 échantillons). Pour ce modèle, 13 variables latentes ont été retenues par validation croisée. Sur la base d'apprentissage, 27 points, soit 77% de la base sont prédits dans l'IC de la mesure de référence (±1,4) et une RMSEC de 1,6 a été obtenue. Le modèle a également été évalué sur la base de test. Seul le coefficient de corrélation obtenu (0,95) a été donné. Ce critère n'est cependant pas suffisant pour évaluer la qualité de la prévision.

Une étude similaire a été réalisée par Braga *et al.* [120] sur une base de données conséquente de 701 lubrifiants de provenance diverse (32 <VI < 182) a été analysée. 473 échantillons ont été utilisés pour la phase d'apprentissage et 231 pour tester le modèle. Deux régions spectrales ont été sélectionnées : la première est comprise entre 3330,1 et 1319 cm^{-1} et la seconde entre 985,0 et 649,9 cm^{-1}. Des RMSEC et RMSEP de 5,2 et 7,1 ont été obtenue sur la base d'apprentissage et sur la base de test respectivement. Ces valeurs sont relativement élevées. De plus, cet écart de deux points entre les deux bases jette un doute sur la capacité du modèle à se généraliser à des échantillons de provenances diverses. Récemment, un modèle de régression PLS a été développé au sein d'IFPEN pour la prédiction du VI de la coupe huile à partir de données ^{13}C RMN du liqtot afin de répondre à la problématique posée par l'utilisation d'EHD [15]. Les caractéristiques du modèle sont récapitulées dans le Tableau 8. Sur une base d'apprentissage constituée de 160 échantillons de VI compris entre 9 et 130, 66% sont prédits avec une erreur inférieure à l'IC de la mesure de référence et une RMSEC de 2,23 a été obtenue. Sur la base de test constituée de 60 échantillons, la RMSEP obtenue est de 4,09. La variété des huiles prises en compte (issues en partie de l'HDT et en partie d'HCK des DSV) témoigne de l'intérêt de cette étude, bien que la précision du modèle ne soit pas tout à fait satisfaisante.

Tableau 8 : Caractéristiques du modèle PLS pour la prédiction du VI à partir de données ^{13}C RMN du liqtot sur la base d'apprentissage

Base de données	Nombre d'échantillons	Plage de VI	Nombre de variables latentes	RMSECV	RMSE*
Apprentissage	160	9 - 130	12	3,4	2,2
Test	62	9 - 130	-	-	4,1

*→RMSEC pour la base d'apprentissage et RMSEP pour la base de test

2.3.6 Prédiction dans le simulateur IFPEN

Pour mieux comprendre les objectifs de la modélisation du PT de la coupe gazole et du VI de la coupe huile, nous revenons ci-dessous sur les contraintes liées au développement de modèle dans le simulateur IFPEN. Comme nous l'avons évoqué en introduction, les propriétés y sont prédites uniquement à partir de propriétés globales des différentes coupes pétrolières (charges, effluents) mises en jeu durant le procédé d'HCK.

2.3.6.1 Cas du PT de la coupe gazole

Nous avons schématisé le simulateur IFPEN dans le cas de la prédiction du PT de la coupe gazole (Figure 9). Ce modèle résulte de deux sous-modèles imbriqués :

1. un premier sous-modèle qui permet d'estimer des propriétés plus accessibles de la coupe gazole (densité, indice de réfraction, températures de distillation simulée, proportion de carbone paraffinique, *etc.*) *via* des équations cinétiques et des corrélations polynomiales (méthode dite n-d-M par exemple [22]) ;
2. un second sous-modèle qui a pour descripteurs des propriétés estimées en (1) et qui les relient au PT de la coupe gazole par régression polynomiale (Eq.1. 5).

Chaque sous-modèle est construit à partir de mesures faites sur les coupes mises en jeu. Cela permet notamment d'étudier leurs performances indépendamment l'un de l'autre, ainsi que la propagation de l'erreur entre l'entrée et la sortie du modèle final. Le sous-modèle (1) donne fournit des performances en adéquation avec les objectifs visées par IFPEN et ne sont donc pas remis en question. **Pour cette thèse, nous nous intéresserons à l'amélioration du sous-modèle (2) reliant le PT aux autres propriétés de la coupe gazole.**

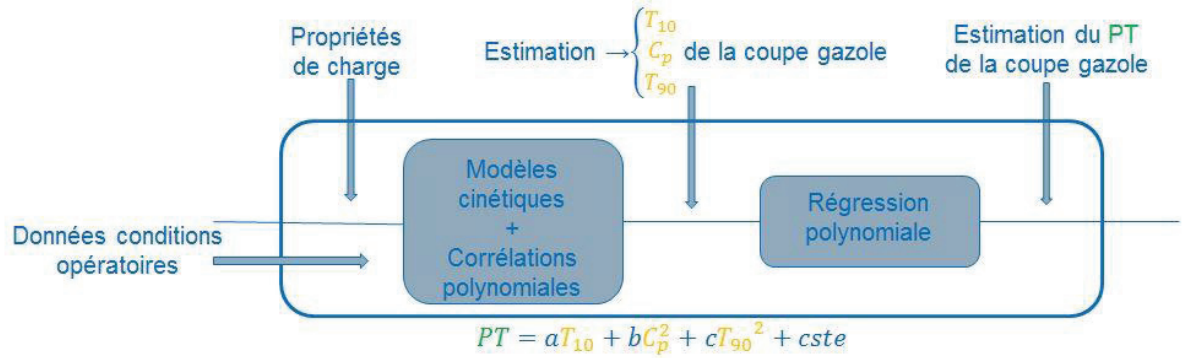


Figure 9 : Schéma simplifié du modèle de prédiction du PT de la coupe gazole dans le simulateur IFPEN

2.3.6.2 Cas du VI de la coupe huile

Bien que l'objectif principal soit la modélisation du VI des huiles produites par HCK, estimer le VI des huiles obtenues par distillation de l'effluent hydrotraité est essentielle. En effet, cette grandeur donne une indication sur la qualité de l'HDT (elle est notamment très reliée à la teneur en aromatiques). Elle est donc systématiquement prédite dans le simulateur IFPEN. Nous avons schématisé ce dernier dans le cas de la prédiction du VI de la coupe huile (Figure 10). Ce schéma est valable aussi bien pour les huiles issues d'HDT que pour celles qui proviennent d'HCK (Eq.1. 15). Dans chaque cas, la charge désigne la coupe pétrolière en entrée du procédé (DSV pour l'HDT et effluent hydrotraité pour l'HCK).

Ces modèles ont également été construits à partir de mesures faites sur les coupes pétrolières mises en jeu. **Dans les deux cas, les performances sont limitées.** Nous nous intéresserons donc à leur amélioration avec la volonté de conserver la structure globale du simulateur IFPEN.

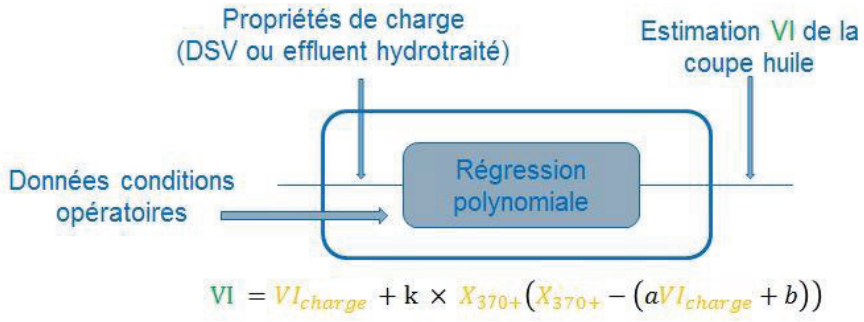


Figure 10 : Schéma simplifié du modèle de prédiction du VI de la coupe huile dans le simulateur IFPEN

2.4 Conclusions sur l'état de l'art

L'étude bibliographique a permis de mettre en évidence la complexité des paramètres qui influent sur le PT de la coupe gazole et le VI de la coupe huile. Elle a également démontré le potentiel de techniques de caractérisation d'échantillon telles que l'IR et la RMN du ^{13}C qui, couplées à des méthodes d'analyse statistique de données fournissent des résultats intéressants.

Concernant la coupe gazole, les études qui ont été recensées ont mis en évidence :

1. **l'impact de la cristallisation des n-paraffines en milieux hydrocarbures sur les propriétés à froid ;**
2. **l'influence du milieu à l'échelle moléculaire** sur le processus de cristallisation de ces n-paraffines ;
3. **l'importance des points de coupe ou intervalles de distillation** (plus ceux-ci sont élevés, plus les propriétés à froid augmentent).

Toutefois l'implication de l'ensemble des n-paraffines de masse moléculaire variable sur la variation du PT n'est pas clairement décrite. De même, l'influence des autres entités moléculaires (isoparaffines, naphtènes, aromatiques) n'est pas clairement définie.

Concernant le VI de la coupe huile, il ressort que **la présence d'isoparaffines monobranchées et de mononaphtènes avec de longues chaînes alkyles engendre une augmentation du VI. Au contraire, la présence de composés aromatiques tend à diminuer le VI.** Par contre, des interrogations subsistent autour de l'influence sur le VI :

- des isoparaffines multibranchées,
- de la faible quantité de n-paraffines encore présentes dans la coupe huile après déparaffinage.

La plupart des études qui ont été recensées sont basées sur un nombre d'échantillons est en général limité. De ce fait, l'absence d'étapes de test permettant de vérifier la consistance des modèles proposés est souvent remarquée. Dans le cas du VI de la coupe huile comme dans celui du PT de la coupe gazole, **les études les plus abouties font appel à des méthodes chimiométriques telles que la régression PLS appliquées à des données spectrales ou chromatographiques.** Toutefois, exceptés dans les travaux de Lacoue-Nègre [15] à partir de la RMN du ^{13}C, les analyses sont faites directement sur des échantillons de produits (coupe gazole ou coupe huile). Les modèles obtenus ne permettent donc pas de répondre à la problématique liée à l'utilisation d'EHD mais ouvrent des pistes intéressantes.

Les modèles de régression polynomiale actuellement implémentés dans le simulateur montrent clairement des limites pour la prédiction des propriétés d'intérêt. **De plus, aucune alternative n'a été identifiée dans la littérature qui réponde à la contrainte d'accessibilité pour le raffineur.**

2.5 Méthodologie de la thèse

Au regard du constat sur l'état de l'art, nous proposons pour les travaux de cette thèse une méthodologie basée sur deux points complémentaires essentiels (Figure 11) :

1. la caractérisation des coupes gazole et huile par des méthodes d'analyse détaillée d'échantillons,
2. l'utilisation de méthodes statistiques pour exploiter les données analytiques dans le but d'expliquer et de prédire les propriétés d'intérêt de ces coupes.

2.5.1 Caractérisation des coupes pétrolières

L'identification des différents composés présents dans les gazoles et les huiles, et notamment la distinction entre les molécules qui ont des structures proches (n-paraffines et isoparaffines par exemple) est primordiale pour la compréhension moléculaire des propriétés d'intérêt. Nous avons également rappelé que le PT de la coupe gazole et le VI de la coupe huile sont liés à la structure moléculaire des composés. Nous proposons donc :

1. la mise en œuvre de la chromatographie en phase gazeuse bidimensionnelle (GC×GC) couplée à un détecteur FID (*Flame Ionisation Detector*) d'hydrocarbures pour obtenir une cartographie détaillée des composés présents dans les coupes ;
2. l'utilisation de la RMN du ^{13}C pour obtenir des informations sur la structure des molécules.

2.5.2 Compréhension moléculaire

L'identification des principaux marqueurs moléculaires qui influent sur une propriété d'une coupe pétrolière peut s'avérer complexe étant donnée la diversité des molécules présentes dans ces coupes. Le nombre de marqueurs peut augmenter fortement (des centaines ou des milliers) lorsque des techniques de caractérisation chromatographiques ou spectroscopiques sont utilisées. Pour exploiter nos bases de données, nous proposons la mise en œuvre successive :

1. d'une de régression PLS sur les données analytiques (GC×GC ou ^{13}C RMN) pour vérifier que l'information nécessaire pour prédire la propriété est bien présente dans ces données,
2. d'une *sparse* PLS (PLS parcimonieuse) pour déterminer les parties intéressantes du signal et améliorer la robustesse des modèles construits.

Cette approche a un double intérêt :

1. l'identification des principaux marqueurs qui influent sur la propriété modélisée,
2. discuter quant à la possibilité de prédire cette propriété à partir du même type de données (GC×GC ou ^{13}C RMN) provenant d'analyse du liqtot.

2.5.3 Prédiction

Etant donnée la restriction autour du type de descripteurs qui peuvent potentiellement être intégrés dans le simulateur IFPEN, l'amélioration des modèles ne peut se faire que sur deux points :

1. l'identification de descripteurs plus pertinents parmi les propriétés globales à disposition
2. l'introduction de modèle plus adapté à la complexité des propriétés à modéliser

Pour le premier point, nous proposons la stratégie suivante :

- présélection de potentiel descripteurs par combinaison entre avis d'expert en procédé (« *brainstorming* ») et résultats de la compréhension moléculaire,
- sélection des descripteurs finaux parmi les présélectionnés par méthode statistique de sélection de variable.

Pour le second point, nous préconisons l'utilisation de méthodes d'interpolation (*splines* ou krigeage) qui présentent les avantages suivants :

- une adaptation automatique à la prédiction de propriétés linéaires ou non linéaires
- une bonne gestion des données complexes (présence de clusters, irrégularité de la propriété à modéliser)
- une efficacité relative dans le cas où on dispose d'un nombre d'échantillon limité.

Notons que dans le cas de la coupe huile, nous nous imposerons le challenge de prédire le VI à partir de propriétés globales du liqtot. Ce choix est motivé par deux facteurs :

1. les résultats encourageants des travaux de Lacoue-Nègre [15] qui mettent en avant cette possibilité ;
2. la perspective de pouvoir estimer le VI à partir du liqtot plus aisément que *via* des données spectroscopiques.

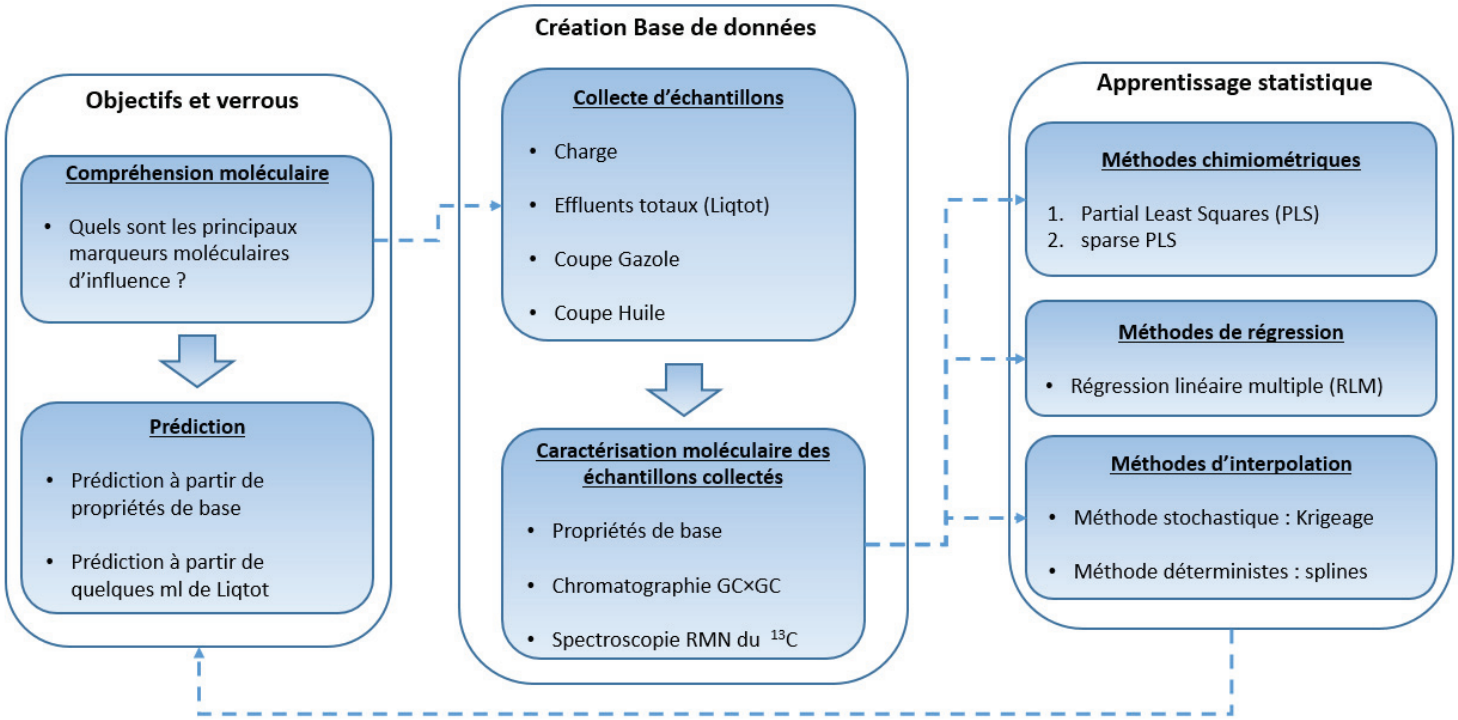


Figure 11 : Objectifs, verrous et méthodologie de la thèse

Partie 2 : Matériel et Méthodes

L'objet de cette seconde partie est de présenter les différentes méthodes de caractérisation des coupes pétrolières et d'analyse de données qui ont été utilisées au cours de cette thèse. Nous avons évoqué deux formes de caractérisation : la caractérisation moléculaire issue de méthode d'analyse physico-chimiques (^{13}C RMN et GC×GC) et la caractérisation globale par des propriétés mesurées à partir de méthodes standardisées. De même, pour chaque type de données nous avons choisi de recourir à des méthodes statistiques adaptées aux objectifs fixés : d'une part les méthodes chimiométriques type PLS et *sparse* PLS, d'autre part les méthodes d'interpolation telles que le krigeage.

Les 3 prochains chapitres sont organisés comme suit : dans le Chapitre 3 nous décrivons les différentes méthodes d'analyse physico-chimiques qui ont été utilisées pour caractériser les échantillons de coupes. Le Chapitre 4 revient sur l'aspect théorique des modèles qui ont été développées au cours de cette thèse. Enfin, dans le Chapitre 5 nous proposons une présentation globale des données qui ont été recueillies et exploitées pour cette étude.

Chapitre 3. Méthodes analytiques

Dans ce chapitre nous présentons les différentes méthodes analytiques qui ont été utilisées au cours de cette thèse pour la caractérisation des différentes coupes pétrolières : charges, effluents, coupe gazole et coupe huile. Outre les propriétés d'usage dont les méthodes standards de mesures sont rappelées, ainsi que certaines corrélations connues dans la littérature [90], nous revenons sur les principes de fonctionnement de la GC×GC et de la RMN du ^{13}C.

Les premiers exemples de chromatographie multidimensionnelle remontent à la chromatographie planaire [121] avec l'élution de deux phases mobiles suivant des directions orthogonales. Mais c'est l'introduction par Giddings [122] du concept de méthodes multidimensionnelles qui a permis de réellement apprécier les nouvelles perspectives les sciences séparatives. A la suite de ces travaux, plusieurs combinaisons de techniques de séparation ont été développées : LC×LC [37,39], LC×GC [123,124] ou encore SFC×GC [125]. Depuis son introduction par Phillips *et al*. [126] en 1991, la GC×GC a connu un franc succès. De nombreuses études scientifiques [127–130] y ont d'ailleurs été consacrées. Au sein d'IFPEN, plusieurs travaux ont été menés autour de cette technique pour la caractérisation des coupes pétrolières [35,131,132] des fractions légères (kérosène, gazole, *etc.*) aux fractions lourdes (DSV, effluents hydrotraités).

Depuis l'utilisation en 1958 par William et *al*. [133] de la ^{1}H RMN pour la caractérisation des hydrocarbures dans les pétroles, l'utilisation de la spectroscopie RMN comme une technique d'identification de structure moléculaire est devenue un challenge pour les chimistes et les physiciens [32]. Comme le mentionne Edwards [34], la littérature associée à ce type de travaux s'étend presque sur 60 ans et est difficile à synthétiser en un chapitre. Dans le domaine des produits pétroliers, la RMN du ^{13}C a été de plus en plus plébiscitée au cours des dernières décennies, principalement en raison d'une meilleure résolution des spectres ^{13}C RMN en comparaison de la ^{1}H RMN. De nombreux travaux ont depuis été publiés autour de la caractérisation des coupes pétrolières par ^{13}C RMN [71,107–109,112–114,117,118,134]. Bien que des limites aient été identifiées, notamment autour de la quantification des différents types de carbone (primaire, secondaire, tertiaire ou quaternaire), donnant ainsi lieu à l'émergence de la RMN multi-impulsionnelle [113] ou de la RMN 2D [93,135,136], l'exploitation de spectre ^{13}C RMN par des techniques d'analyses de données multivariées donne encore des résultats probants [73].

Nous présenterons dans un premier temps l'ensemble des propriétés d'usage qui sont régulièrement mesurées sur les coupes pétrolières et mise à disposition des raffineurs. Nous reviendrons ensuite sur chacune des méthodes de caractérisation moléculaire (GC×GC et ^{13}C RMN) qui ont été mise en œuvre au cours de cette thèse. Pour chaque méthode nous décrirons son principe de fonctionnement.

Le mode d'acquisition, le prétraitement et la mise en forme des données avant exploitation seront également détaillés.

3.1 Mesures des propriétés d'usage

Plusieurs propriétés physico-chimiques dites d'usage sont couramment utilisées en raffinerie pour caractériser globalement les différentes coupes pétrolières [3,90]. Certaines d'entre elles sont mesurées directement à partir d'un échantillon de la coupe suivant des essais normalisés. Les autres sont déduites à partir de corrélations mathématiques connues dans la littérature.

3.1.1 Analyse élémentaire des coupes pétrolières

L'objet de cette analyse est de déterminer la composition élémentaire des coupes pétrolières. Cette étape est primordiale pour situer la qualité d'une coupe ou l'efficacité d'un traitement de raffinage [3]. Elle consiste en grande partie à la détermination de la concentration en hydrogène et carbone d'une part, puis en azote, en soufre et en métaux (qui sont plutôt caractéristiques de produits de mauvaise qualité).

3.1.2 Distillation simulée (DS)

Bien que la distillation atmosphérique (ASTM D86 [4]) soit une méthode très pratique et relativement simple à mettre en œuvre, elle manque de robustesse et a une mauvaise reproductibilité. De plus elle requiert une quantité importante de produit (environ 100 ml). Une autre méthode de caractérisation par GC (DS) a ainsi été introduite et est fortement préconisée. La méthode est décrite par la norme ASTM D2887 [98]. Elle fournit notamment une série de températures d'ébullition caractéristique de l'échantillon analysé. Par la suite, nous noterons T_5, T_{10}, T_{15}, …, T_{95} les points d'ébullition issus de la DS, où T_i désigne la température pour laquelle i % de la masse du produit a été distillée. Dans le cas de mélanges complexes, la notion de point d'ébullition moyen ou *Mean Average Boiling Point* ($MeABP$) est également utilisée. Une définition précise de cette grandeur est donnée par Riazi [90]. Bien qu'elles soient rigoureusement différentes, la $MeABP$ est souvent prise en première approximation égale (à un facteur additif près) à la T_{50}.

3.1.3 Propriétés physico-chimiques globales

3.1.3.1 Masse volumique et densité

En toute rigueur, la masse volumique est définie comme la masse par unité de volume d'un fluide. Dans le cas des produits pétroliers liquides, elle est usuellement reportée en termes de densité relative notée $d_{t_2}^{t_1}$:

$$\boldsymbol{d_{t_2}^{t_1} = \frac{masse\ volumique\ du\ liquide\ \text{à}\ t_1°C}{mase\ volumique\ de\ l'eau\ \text{à}\ la\ température\ t_2°C}} \qquad \textbf{(Eq. 2. 1)}$$

La grandeur la plus couramment utilisée est soit la d_4^{15} (la masse volumique de l'eau à 4°C est exactement égale à 1 g/cm^3) ce qui permet de relier plus aisément cette grandeur à la masse volumique du fluide), soit la d_{15}^{15} aussi connue sous le nom de *specific gravity* ($Spgr$).

3.1.3.2 Masse molaire

La masse molaire (M) est une propriété importante pour la caractérisation des coupes pétrolières. Elle fournit des indications importantes sur la taille et la structure des molécules. La masse molaire d'une coupe pétrolière est définie comme la valeur moyenne des masses molaires de ses constituants :

$$\boldsymbol{M = \sum_i x_i M_i} \qquad \textbf{(Eq. 2. 2)}$$

x_i et M_i sont respectivement la fraction et la masse molaire du composé i. Les méthodes de mesure de M dans les fractions pétrolières sont particulièrement peu fiables. Elle est donc généralement déterminée à partir de corrélations [90].

3.1.3.3 Indice de réfraction

L'indice de réfraction (n) pour un fluide donné est le ratio de la vitesse de la lumière dans le vide sur la vitesse de la lumière dans le fluide à une température donnée. Sa méthode de mesure est donnée par la norme ASTM D1218 [137]. Comme la densité, l'indice de réfraction des hydrocarbures varie à la fois avec leur structure chimique (IR paraffines < IR naphtènes < IR aromatiques) et croît avec la masse molaire [3].

3.1.3.4 Viscosité dynamique et viscosité cinématique

La viscosité est la mesure du frottement fluide d'une couche sur une autre. Si on isole une couche d'un fluide en mouvement, on observe un caractère visqueux lorsque l'on distingue le mouvement relatif de la couche isolée par rapport à ses couches voisines. Elle constitue la propriété essentielle d'un lubrifiant. Il existe deux mesures de la viscosité : la viscosité dynamique (η) qui caractérise le mouvement relatif de deux couches voisines ; la viscosité cinématique (ν) définie comme le rapport de la viscosité dynamique sur la masse volumique (ρ). C'est cette dernière qui est couramment utilisée dans le domaine pétrolier. La viscosité cinématique s'exprime en centistockes (cSt) [90]. Elle est obtenue en mesurant le temps d'écoulement de l'huile dans un tube capillaire en verre calibré suivant la norme ASTM D445 [138] ou la norme ASTM D7042 [26] en déterminant d'abord η et ρ du produit à l'aide d'un viscomètre équipé d'un tube oscillant en U. On en déduit alors ν par la relation :

$$\boldsymbol{\nu = \frac{\eta}{\rho}} \qquad \textbf{(Eq. 2. 3)}$$

La viscosité varie avec la pression mais l'influence de la température est beaucoup plus marquée [3]. En effet, la viscosité diminue rapidement avec cette dernière. De nombreuses équations et abaques relient

la viscosité à la température, notamment la norme ASTM D341 [23]. Cette dernière permet d'évaluer la viscosité des produits pétroliers liquides à une température donnée.

3.1.3.5 Autres propriétés

Des corrélations utilisant les données de la DS, n, d et M ont été établies, qui permettent d'estimer d'autres propriétés de produits pétroliers. Certaines d'entre elles sont données dans le Tableau 9**.**

Les pourcentages en carbones paraffiniques, naphténiques et aromatiques sont par exemple estimés par méthode dite n-d-M [22]. **Notez qu'un carbone commun à deux structures (aromatique et naphténique) ou (aromatique et paraffinique) ou encore (naphténique et paraffinique) sera d'abord aromatique, puis naphténique, puis paraffinique.**

Le facteur de caractérisation de Watson (K_w) est une grandeur physique qui a été introduite par les chercheurs de la Société « Universal Oil Products Co » [3]. Il a été construit sur le double constat que la densité des hydrocarbures dépend du rapport H/C et que leur point d'ébullition est lié à leur nombre d'atomes de carbone.

Les méthodes de mesure des propriétés à froid de la coupe gazole (PE, PT et TLF) et du VI de la coupe huile ont déjà été introduites aux paragraphes 2.2 et 2.3 respectivement.

Tableau 9 : Propriétés de base pour la caractérisation des coupes pétrolières

Propriétés	Méthodes standards ASTM/NF ISO/IFP	Références
% m/m en Hydrogène	ASTM D5291	[139]
% m/m en Carbone	ASTM D5291	[139]
% m/m en Soufre	IFP 9910	
% m/m en Azote	IFP 9608	
Densité (d)	NF EN ISO 12185	[17]
Indice de réfraction (IR)	ASTM D1218	[137]
Distillation simulée ($T_i, i = 5, 10, 15, \dots, 90, 95$)	ASTM D2887	[98]
Caractérisation par type de carbone (Ca,Cp,Cn)	ASTM D3238 ou méthode n-d-M	[22]
Nombre moyen de cycles aromatiques (Ra)	ASTM D3238	[22]
Masse molaire (M)	DT – 50	Basé sur [90]
Mean Average Boiling Point ($MeABP$)	Corrélation API	[90]
Facteur de caractérisation de Watson (Kw)	$K_w = \frac{(1.8\ MeABP)^{\frac{1}{3}}}{Spgr}$	[90]
Température moyenne (TM)	$TM = \frac{T_{10} + T_{50} + T_{90}}{3}$	[7]
Température moyenne pondérée (TMP)	$TMP = \frac{T_5 + 2T_{50} + 4T_{95}}{7}$	[7]
Gravité spécifique ($Spgr$)	$Spgr = 1.001 d_4^{15}$	[90]
$API\ Gravity$	$API.Gravity = \frac{141.5}{Spgr} - 131.5$	[90]
Point de trouble (PT)	ASTM D2500 / NF ISO EN 23015 (coupe gazole)	[81] / [18]
Point d'écoulement (PE)	ASTM D97 / NF T60-105 (coupe gazole)	[27] / [21]
Température limite de filtrabilité (TLF)	ASTM D6371 / NF EN 116 (coupe gazole)	[25] / [82]
Viscosité cinématique (ν)	ASTM D445 / ASTM D7042	[138] / [26]
VI	ASTM D2270 / NF ISO 2909	[104] / [19]

3.2 Caractérisation moléculaire des coupes pétrolières

La complexité des mélanges qui constituent la coupe gazole et la coupe huile poussent à l'utilisation de différentes techniques de caractérisation moléculaire qui permettent l'identification et la quantification de types de molécules ou de structures moléculaires. Dans notre étude nous avons choisi la GC×GC et la RMN du ^{13}C. Le principe de ces deux techniques et le traitement des données acquises en vue de leur utilisation par des méthodes chimiométriques sont décrits dans ce paragraphe.

3.2.1 La chromatographie en phase gazeuse bidimensionnelle (GC×GC)

3.2.1.1 Principe de fonctionnement

La GC×GC est une technique de séparation qui permet en une injection de caractériser des échantillons complexes [35]. Elle met en jeu deux colonnes GC de polarité différente. Par l'intermédiaire

d'un modulateur, les effluents issus de la première colonne sont échantillonnés pour être réinjectés dans la seconde colonne sous forme de différentes fractions en fonction de la période de modulation définie. De manière analogue à la GC-1D, la GC×GC disposera d'un système d'injection, de détection mais également d'un système de traitement spécifique du signal.

Les effluents étant généralement échantillonnés en continu par le modulateur, la seconde séparation devra donc être très rapide (quelques secondes). Cet aspect est particulièrement important car chaque fraction réinjectée devra être séparée dans un temps inférieur à la durée de l'échantillonnage (période de modulation : P_{Mod}). Généralement, les deux colonnes chromatographiques sont placées dans le même four. La structure d'un chromatogramme 2D est illustrée sur la Figure 12.

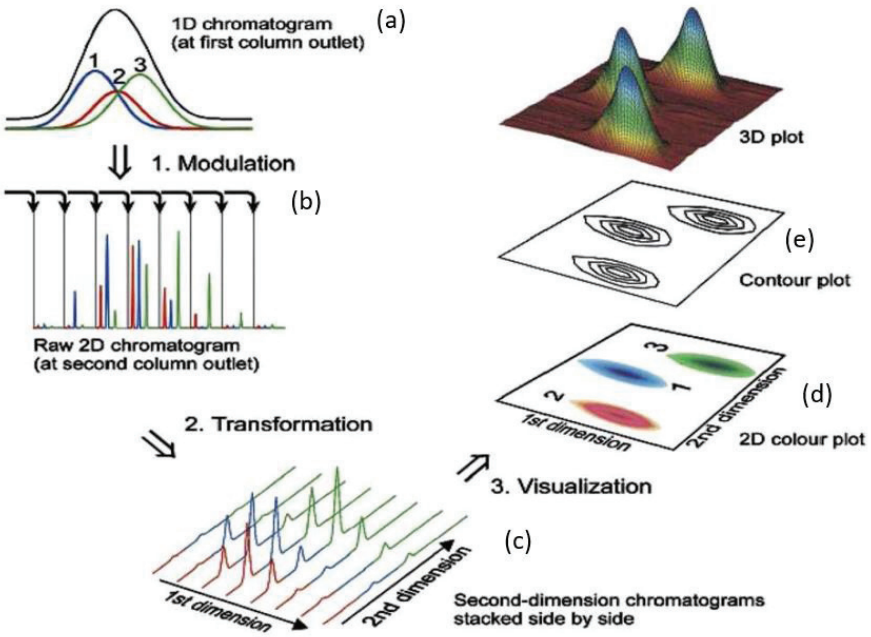


Figure 12 : Principe de retraitement des chromatogrammes GC×GC ; (a) chromatogramme 1D ; (b) chromatogramme modulé ; (c) chromatogrammes empilés ; (d) chromatogramme GC×GC [35]

Après échantillonnage des pics 1D et analyse dans la seconde colonne, le signal acquis par le détecteur est une succession d'analyses de seconde dimension accolées. Le signal est transformé suivant un plan de rétention avec des axes correspondant à chaque dimension de la séparation 2D sous la forme d'un chromatogramme 2D, montrant des nuances d'intensité et des courbes d'iso-intensité.

3.2.1.2 GC×GC-FID

La GC×GC – FID est utilisée pour la caractérisation des échantillons de gazole selon la méthode IFPEN 1602 [140]. Un schéma du dispositif est proposé sur la Figure 13. Le chromatographe 2D est équipé d'un régulateur de pression et d'un couple de colonnes capillaires greffées avec une phase

stationnaire polaire de type polyméthylsylphénylènesiloxane. Dans la première dimension, la colonne (1) (Figure 13) de type apolaire fournit une séparation en fonction des points d'ébullition des molécules alors que dans la deuxième dimension, la colonne (2) (Figure 13) de type polaire sépare les composés suivant leur polarité (soit le type d'hydrocarbure). Les composés sont élués avec l'hélium comme gaz vecteur et détectés avec un détecteur FID. Ainsi, chaque composé élué est caractérisé par un couple de temps de rétention (tr_1, tr_2) correspondant à chacune des dimensions de séparation. Le programme de traitement des données utilise les données brutes (3) (Figure 13) pour la visualisation 2D/3D des chromatogrammes (4) (Figure 13) et pour leur intégration.

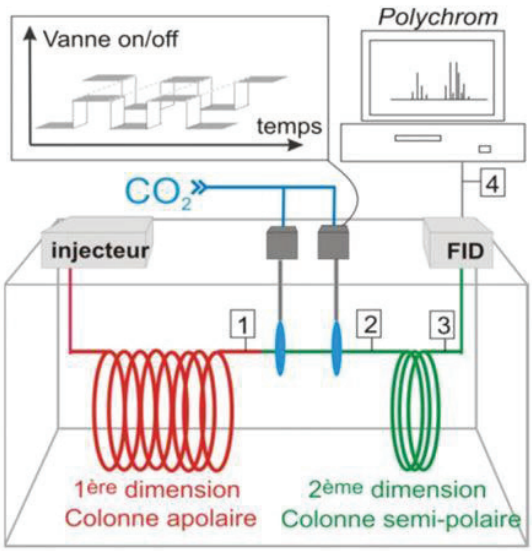


Figure 13 : Schéma simplifié du dispositif GC×GC [140]

La Figure 14 représente le chromatogramme 2D d'un gazole produit par HCK. Les intensités des pics sont mesurées par une échelle de couleur (du blanc au bleu). On peut distinguer les zones du chromatogramme correspondant à différentes familles d'hydrocarbures saturés et aromatiques (Figure 14a). Les paraffines sont les hydrocarbures les moins polaires, suivi des naphtènes puis aromatiques. Pour les composés naphténiques, la polarité augmente avec le nombre de cycles naphténiques. De même, la polarité d'un composé aromatique est une fonction croissante du nombre de cycles benzéniques qu'il contient. Les pics chromatographiques générés par un même type de composés durant le processus de modulation apparaissent dans le chromatogramme 2D comme un ensemble de tâches regroupées en « blobs » (Figure 14b). La concentration dans le mélange de chaque composé ainsi répertorié est proportionnelle à l'aire du blob correspondant [35].

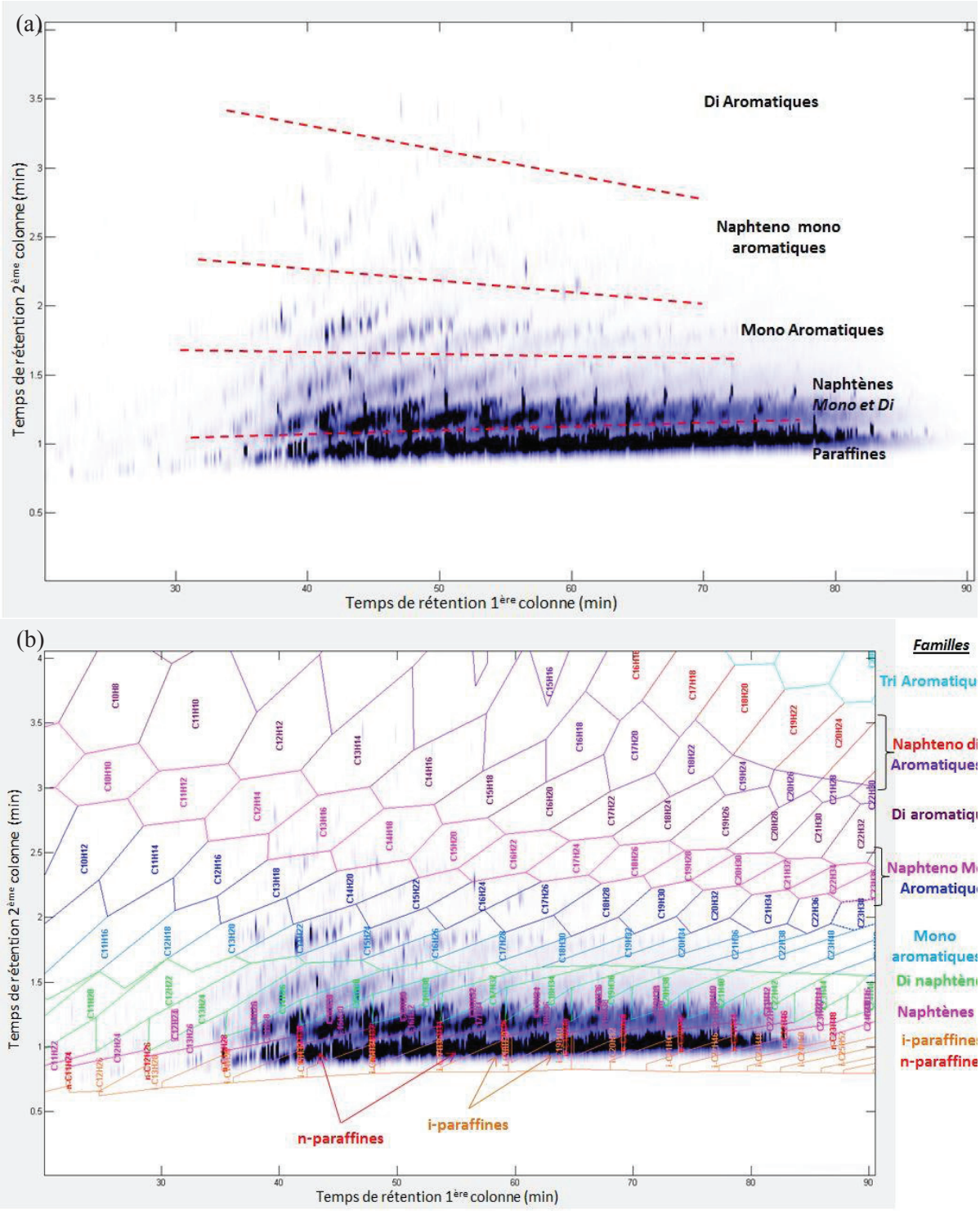


Figure 14 : Chromatogrammes 2D issues de l'analyse GC×GC d'un gazole produit par hydrocraquage ; a) Localisation des différentes familles d'hydrocarbures ; b) Localisation des blobs pour l'identification des différentes espèces

Une fois le masque d'identification recalé pour les différents blobs selon la méthode IFPEN 1602 [140], les aires correspondantes sont calculées et rangés sous la forme d'un tableau tel que représenté sur la Figure 15. Le tableau représente la proportion massique (% m/m) de chaque composé identifié en fonction de son nombre de carbone (lignes) et de sa famille chimique (colonnes). La GC×GC – FID permet d'identifier des composés appartenant 16 familles différentes d'hydrocarbures dans la coupe gazole [35]. Le récapitulatif de ces familles est donné en Annexe F. Par exemple le coefficient situé à ligne 26 de la 17ème colonne est la proportion massique de l'isoparaffine de formule brute $C_{26}H_{54}$ et vaut 0,012 % m/m. **Seules les lignes qui contiennent au moins un coefficient non nul sont prises en compte**. Chaque échantillon est finalement représenté par un vecteur contenant les coefficients du tableau pris ligne par ligne (Figure 16).

.

Formule brute des différentes familles d'hydrocarbures

Nombre de C	CnH2n	CnH2n-10	CnH2n-12	CnH2n-14	CnH2n-16	CnH2n-18	CnH2n-2	CnH2n-20	CnH2n-22	CnH2n-24	CnH2n-26	CnH2n-28	CnH2n-30	CnH2n-4	CnH2n-6	CnH2n-8	i-CnH2n+2	n-CnH2n+2
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0.059	0	0.011	0
8	0.024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0.074	0	0	0	0	0	0.012	0	0	0	0	0	0	0	0.014	0	0.049	0.014
10	0.084	0	0	0	0	0	0.02	0	0	0	0	0	0	0	0.01	0	0.059	0
11	0.056	0	0	0	0	0	0.053	0	0	0	0	0	0	0	0.014	0	0.057	0
12	0.072	0.051	0	0	0	0	0.13	0	0	0	0	0	0	0	0.022	0.029	0.076	0.016
13	0.36	0.15	0.062	0.013	0	0	0.654	0	0	0	0	0	0	0	0.288	0.18	0.201	0.097
14	2.474	0.194	0.118	0.032	0.011	0	2.38	0	0	0	0	0	0	0	1.138	0.344	2.193	0.767
15	4.321	0.138	0.076	0.046	0.016	0	2.665	0	0	0	0	0	0	0	1.284	0.288	5.983	0.945
16	3.55	0.172	0.058	0.022	0.023	0.028	2.364	0	0	0	0	0	0	0	1.051	0.279	6.604	0.774
17	3.152	0.08	0.041	0.024	0.017	0.032	1.708	0.014	0.014	0	0	0	0	0	0.783	0.338	6.712	0.709
18	2.82	0.064	0.034	0.015	0.011	0.012	1.453	0	0	0	0	0	0	0	0.546	0.21	5.29	0.606
19	2.36	0.045	0.031	0	0	0	1.125	0	0	0	0	0	0	0	0.411	0.13	5.333	0.52
20	1.667	0.035	0.012	0	0	0	1.166	0	0	0	0	0	0	0	0.338	0.092	5.119	0.415
21	1.357	0.018	0	0	0	0	0.939	0	0	0	0	0	0	0	0.276	0.053	4.042	0.31
22	0.611	0	0	0	0	0	0.484	0	0	0	0	0	0	0	0.178	0.013	2.66	0.11
23	0.164	0	0	0	0	0	0.131	0	0	0	0	0	0	0	0.096	0	1.23	0.017
24	0.025	0	0	0	0	0	0.012	0	0	0	0	0	0	0	0.02	0	0.293	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.109	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.012	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

%m/m en iso-$C_{26}H_{54}$

Figure 15 : Tableau des données chromatographiques générées par le logiciel 2D Chrom™

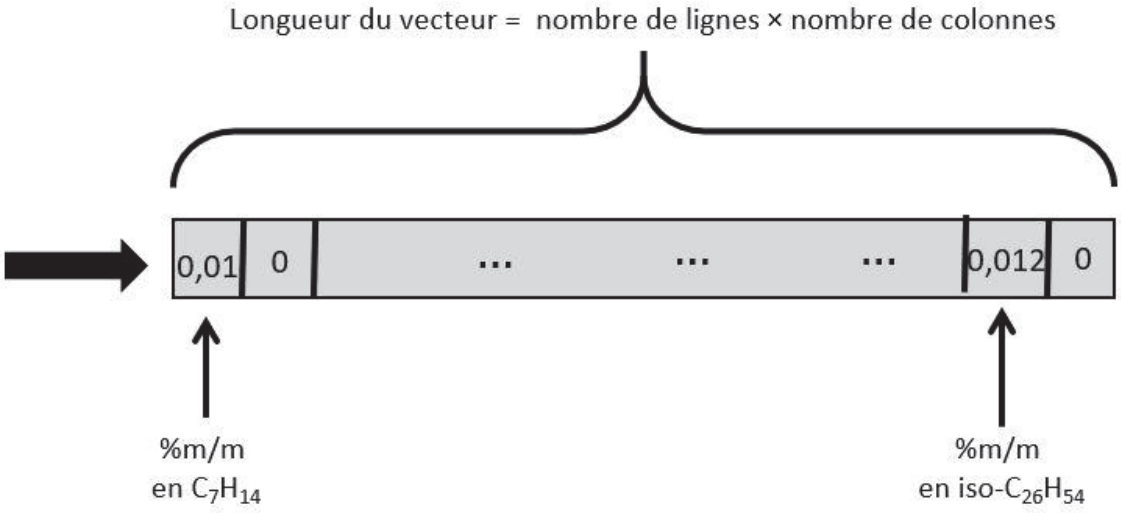


Figure 16 : Mise en forme vectorielle des données GC×GC

3.2.1.3 GC×GC haute température pour l'analyse des huiles

La GC×GC haute température (GC×GC-HT) est utilisée pour la caractérisation des distillats sous vide et des liqtot issus de procédé d'HCK [140]. Un schéma du dispositif est donné sur la Figure 17. Dans ce cas, le chromatographe est équipé d'un injecteur *on-column*, d'un système *backflush*, d'un système de modulation à flux, d'un couple de colonnes capillaires greffées, de différents capillaires inertes et de deux détecteurs à ionisation de flamme. La phase stationnaire apolaire de la première colonne est de type polydiméthylsiloxane et la phase stationnaire polaire de la deuxième colonne est de type phényle / diméthylpolysiloxane. Cette technique permet la quantification des hydrocarbures par famille de composés qui ont entre 8 et 45 atomes de carbone c'est-à-dire pour des points d'ébullition compris entre 126°C et 550°C [140]. Pour chaque famille d'hydrocarbure la DS est calculée.

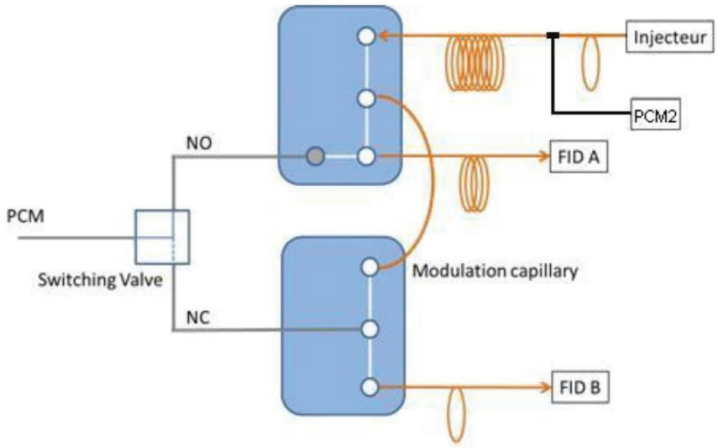


Figure 17 : Schéma de principe de la GC×GC-HT [140]

3.2.1.1 Acquisition et exploitation des données chromatographiques dans le cas des huiles

Dans le cas de la GC×GC – HT, la résolution des signaux 2D ne permet pas de distinguer directement les composés appartenant à une même famille d'hydrocarbures à cause de la complexité de l'échantillon d'huile. Pour chaque famille identifiée, on a accès :

- à la concentration globale de molécules appartenant à cette famille
- aux données de DS [98] de la famille considérée.

Dans le cas des huiles, on distingue 6 familles différentes : les n-paraffines, les iso-paraffines, les naphtènes (mono ou polycycliques), les monoaromatiques, les diaromatiques et les triaromatiques ou autres (tétraaromatiques, polynaphténoaromatiques, *etc*). Pour obtenir une estimation de la concentration des composés du mélange, un découpage de la DS a été effectué pour chaque famille en prenant les températures d'ébullition des n-paraffines comme référence. La procédure pour une famille donnée est la suivante :

- pour chaque température T_i de la DS
  - on recherche le couple de températures de référence consécutives $(T_{réf,1}, T_{réf,2})$ tel que $T_{réf,1} \leq T_i < T_{réf,2}$
  - on attribue la proportion massique évaporée entre T_i et T_{i+1} (ici 0,5%) aux composés de la famille considérée de même nombre de carbones que la n-paraffine de température d'ébullition égale à $T_{réf,1}$.

Pour un type de composé donné, la proportion massique estimée est la somme des proportions massiques qui lui ont été attribuées. A l'issue de cette procédure, un tableau analogue à celui qui a été présenté dans le cas des gazoles est obtenu, chaque colonne faisant référence aux familles répertoriées dans les huiles et chaque ligne renvoyant à un nombre de carbones. La mise en forme des données en vue d'une analyse multivariée est identique à celle qui a été présentée sur la Figure 15.

3.2.1.2 Appareillage

Les analyses GC×GC – FID des échantillons de coupe gazole et GC×GC – HT des échantillons de coupe huile ont été effectuées au département de Caractérisation des Produits (R052) d'IFP Energies Nouvelles à Solaize, France. L'acquisition des chromatogramme 2D a été faite sur des chromatographes GC Agilent 7890. Le modulateur utilisé est un modulateur différentiel à flux (Agilent, modifié d'après les travaux de Griffith *et al.* [141]. Les conditions opératoires sont précisées dans le Tableau 10.

L'acquisition des données a été effectuée avec le logiciel 2D Chrom™ et leur exploitation avec le logiciel statistiques R.

Tableau 10 : Conditions opératoires de la GC×GC – FID pour l'analyse des coupes gazole et la GC×GC – HT pour l'analyse des coupes huile

GC×GC - FID	Injection	Volume / Ratio de split : 0,5 µl / 1/200
	Gaz vecteur Hélium	Débit constant : 1 ml/min (en sortie de 2ème colonne)
	Jeu de colonnes	1ère : PONA[a] 20 m × 0,02 mm ×0,5 µm 2nde : BPX-50[b] 0,8 m × 0,1 mm × 0,1 µm Température initiale : 60°C Programmation : 2°C/min Température finale : 350°C
GC×GC - HT	Injection	Volume / Ratio de split : 0,5 µl / 1/200
	Gaz vecteur Hélium	Débit constant : 1 ml/min (en sortie de 2ème colonne)
	Jeu de colonnes	1ère : 1. [c] 5m × 0,1 mm × 0,4 µm 2ème : 2. [d] 5m × 0,25 mm × 0,1 µm Température initiale : 35°C Programmation : 2°C/min Température finale : 350°C

[a]polydiméthylsiloxane ; [b]polyméthylsylphénylènesiloxane ; [c] diméthylpolysiloxane ; [d]35% phényle/65% diméthylpolysiloxane

3.2.2 La Résonance Magnétique Nucléaire du ^{13}C

3.2.2.1 Principe

La RMN est une technique de caractérisation d'échantillon qui consiste à observer la réponse d'un certain type de noyaux atomiques (soumis à un champ magnétique intense) à un champ magnétique dit de « radiofréquence ». Cette méthode ne peut s'appliquer qu'à des atomes dont le noyau possède un

spin non nul. La RMN du ^{13}C concerne l'isotope 13 du carbone qui possède un spin nucléaire de ½. Soumis à un champ magnétique $\overrightarrow{B_0}$, les noyaux atomiques sont assimilables à des aimants caractérisés par un moment magnétique μ, proportionnel au spin magnétique du noyau et orienté parallèlement à $\overrightarrow{B_0}$ ((a), Figure 18). Par la suite, les noyaux sont soumis à une perturbation sous la forme d'un champ magnétique radiofréquence $\overrightarrow{B_1}$, perpendiculaire à $\overrightarrow{B_0}$; ce qui a pour conséquence un basculement de l'aimantation dans le même plan que $\overrightarrow{B_1}$ ((b), Figure 18). Aussitôt après l'impulsion, on étudie le retour à l'état stable du système excité, suivant un mouvement de précession autour de $\overrightarrow{B_0}$ à une fréquence dite fréquence de Larmor (ν). On peut ainsi mesurer suivant l'axe y (direction de $\overrightarrow{B_1}$) la variation de l'aimantation en fonction du temps ((c), Figure 18). Le signal obtenu communément appelé *FID* (pour *Free Induction Decay*) est donc une superposition de sinusoïdes décroissantes en fonction du temps ((d), Figure 18), chaque sinusoïde correspondant au processus de relaxation caractéristique de chaque noyau et de son environnement. Le spectre RMN est finalement obtenu par application d'une transformée de Fourier au signal *FID*.

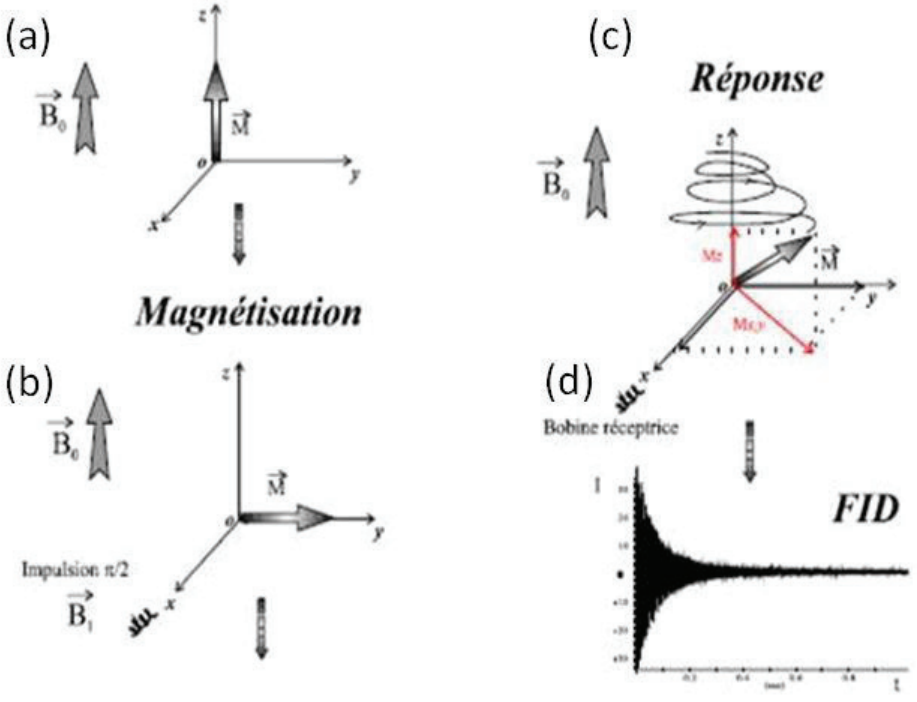


Figure 18 : Schéma de principe de la RMN [15]

Le champ magnétique local peut différer du champ appliqué $\overrightarrow{B_0}$ en particulier du fait du nuage électronique local entourant le noyau qui peut induire un effet perturbateur. De ce fait, le champ magnétique effectivement ressenti par les noyaux est légèrement différent de $\overrightarrow{B_0}$. Cela induit une différence de fréquence de résonance des noyaux observés par rapport à la fréquence de Larmor, appelée déplacement chimique et noté δ. Ce déplacement chimique est donc fonction de l'environnement local (électronique et chimique) du noyau observé et donne donc accès à la coordinence du noyau et à la nature de ses voisins. En pratique, le déplacement chimique est repéré par rapport à la fréquence d'un composé de référence. C'est ce qui permet aussi de comparer les déplacements chimiques de produits connus, quel que soit le champ magnétique utilisé :

$$\boldsymbol{\delta = \frac{\nu_{induit} - \nu_{réf}}{\nu_0}} \qquad \textbf{(Eq. 2. 4)}$$

Où ν_{induit} est la fréquence induite par l'environnement local du noyau et $\nu_{réf}$ est la fréquence du composé de référence.

3.2.2.2 Appareillage

L'acquisition des spectres RMN du ^{13}C a été réalisée sur un spectromètre Bruker Advance 600 MHz (14,1 T). Les échantillons analysés ont été dilués dans du $CDCl_3$ à iso proportion (50/50) en volume. La séquence de découplage « *inverse gated* » [33], qui permet d'effectuer des analyses quantitatives et de comparer des intensités de signaux a été utilisée comme suit : angle de pulsation de 90° (4,5 µs) ; temps d'acquisition (0,769 s) ; temps de relaxation (5,0 s) ; 65 K de données et 512 scans ; température de l'échantillon (25°C).

3.2.2.3 Acquisition et traitement des données spectroscopiques

Le spectre RMN du ^{13}C d'un échantillon d'huile issue de l'HCK des DSV est représenté sur la Figure 19. La zone caractéristique des carbones aliphatiques ou saturés (paraffiniques et naphténiques) est située entre 0 et 60 ppm. Celle des carbones aromatiques se situe entre 117 et 150 ppm. Le triplet de pics caractéristiques du solvant $CDCl_3$ est observé autour de 76,9 ppm. Les données numériques provenant du signal ^{13}C RMN sont sous la forme d'une série d'intensité correspondant à une séquence de déplacements chimique. Les déplacements chimiques sont pris à intervalles réguliers avec un pas d'environ 0,004 ppm. Compte tenu des conditions d'acquisition des spectres précisées au paragraphe précédent, cela représente une séquence de 65536 déplacements chimiques pour un même nombre d'intensité relevé.

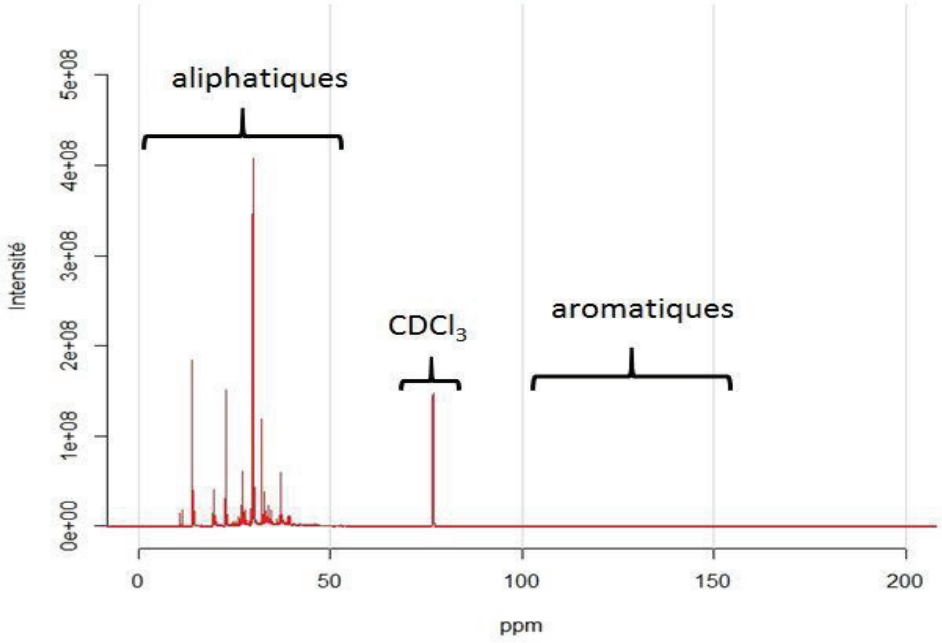


Figure 19 : Spectre RMN du ^{13}C d'un échantillon d'huile issu de l'hydrocraquage des DSV

Les données RMN peuvent être plus ou moins bruitées et la complexité des échantillons peut entraîner des chevauchements de pics importants. Comme toute technique analytique, la détermination du déplacement chimique δ d'un pic caractéristique de l'échantillon est entachée d'une erreur expérimentale, certes faible qui peut gêner les traitements statistiques. De plus, le déplacement chimique, comme nous l'avons mentionné précédemment est lié à la différence de composition moléculaire entre les échantillons d'une même base de données (l'environnement moléculaire d'un noyau atomique influant sur son déplacement chimique). De même la mesure de l'intensité d'un pic caractéristique ou de l'aire de ce pic est sujette à une erreur de mesure. Nous avons donc cherché à estimer l'erreur expérimentale sur le déplacement chimique RMN et sur l'intensité du pic en procédant de la manière suivante pour chaque échantillon collecté :

1. Trois préparations différentes de l'échantillon ont été réalisées (prélèvement et dilution dans le $CDCl_3$.
2. Acquisition des trois spectres ^{13}C RMN correspondant à chaque préparation
3. Estimation de la valeur moyenne du déplacement à partir des trois spectres, puis des valeurs de fidélités associées (répétabilité et intervalle de confiance) suivant la norme NF EN ISO 4259 [20] (Annexe D)
4. Intégration par deux opérateurs différents des pics caractéristiques d'intérêt sur chaque spectre

5. Pour chaque pic estimer la valeur moyenne ainsi que les valeurs de fidélités (répétabilité, reproductibilité, intervalle de confiance de l'aire correspondante suivant la norme NF EN ISO 4259 [20] (Annexe B)

Le décalage pouvant apparaître entre deux spectres sera considéré comme significatif si il est en dehors de l'intervalle de confiance estimé. De même, une différence entre deux aires sera considérée comme significative si les intervalles de confiance associés à ces aires sont totalement distincts.

Les résultats de cette étude seront présentés dans la section 7.3.

3.2.2.4 Recalage des spectres RMN

Des méthodes de *binning* (aussi appelé *bucketing*) sont couramment utilisées pour recaler les spectres [142–145]. Le *binning* est une méthode de traitement de données numériques qui a été introduite pour résoudre les problèmes liés aux décalages entre les spectres d'une même base de données. L'utilisation des méthodes de *binning* est très courante en chimiométrie car la qualité des modèles est conditionnée par un certain nombre d'hypothèses (gaussiennes) faites sur les variables descriptives et qui ne sont pas vérifiées en cas de décalage trop important [146,147]. Le principe du *binning* classique est d'effectuer une réduction des données initiales (généralement des données spectrales) en les regroupant. La méthode classique consiste à diviser chaque spectre de la base de données en plusieurs intervalles (nommés *bins* ou *buckets*) de même taille (généralement comprise entre 0,01 et 0,05 ppm). Les intensités correspondant aux fréquences situées à l'intérieur d'un même intervalle (appelé boîte ou *bin*) sont sommées. La valeur prise par l'intervalle (centré sur le déplacement chimique moyen) correspond à cette somme. La largeur de l'intervalle est choisie de sorte à couvrir la variabilité du déplacement chimique autour des pics, ce qui tend à diminuer le décalage entre les pics des spectres [147]. La méthode utilisée est basée sur les travaux de Sousa *et al.* [143]. Ces derniers ont proposé un nouvel algorithme qui consiste à introduire des *bins* de tailles variables selon la partie du spectre considérée. Il est explicité ci-dessous :

1. Calculer le spectre moyen de la base de données
2. Choisir la taille maximale et la taille minimale d'un *bin*
3. Pour chaque *bin*
    a. Rechercher la taille optimale
    b. Mettre à jour la taille du *bin*
    c. Passer au *bin* suivant

La largeur de l'intervalle est choisie de telle sorte que ses bornes correspondent à des minima locaux des intensités du spectre dans l'intervalle de déplacement chimique considéré. L'avantage de cet algorithme, en comparaison de la méthode de *binning* classique est illustré sur la Figure 25. Cette dernière représente un exemple de découpage d'un spectre RMN dans une zone délimitée d'une part

dans le cas d'un *binning* classique (Figure 25a), d'autre part dans le cas d'un découpage adaptatif optimisé (Figure 25b). Chaque intervalle est délimité par deux droites verticales. Dans le cas classique, les intensités correspondant à un même signal de résonance peuvent être réparties entre plusieurs *bins*, distribuant l'information chimique (Figure 25a) [143]. La Figure 25b montre au contraire comment la considération des minima locaux tend à réduire la taille des *bins* compris entre deux signaux de résonance et permet de résoudre le problème engendré par la méthode classique.

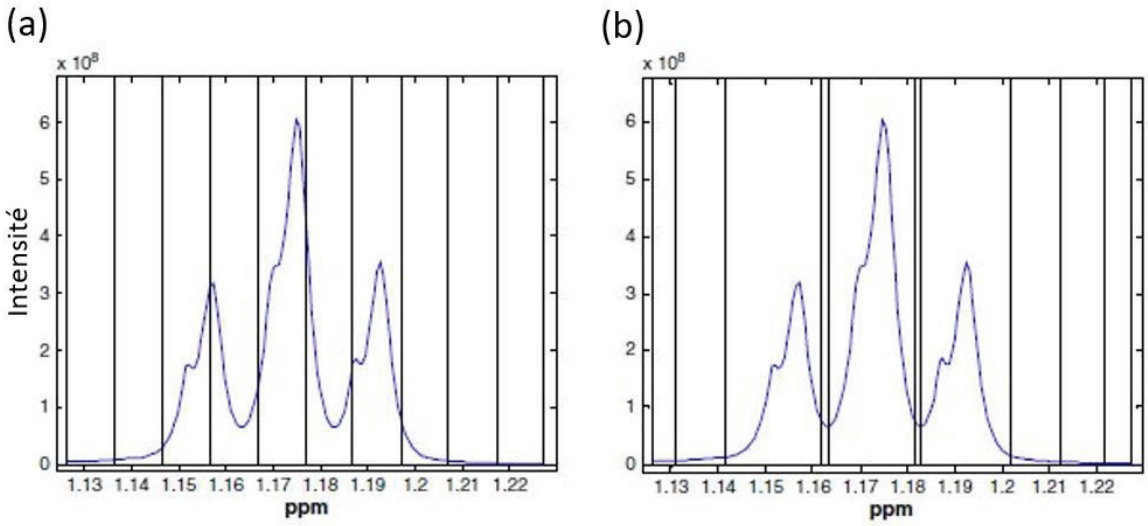


Figure 20 : Illustration du découpage en *bins* d'un spectre RMN ; a) *binning* classique ; b) *binning* adaptatif [143]

3.2.2.5 Méthode de lissage des spectres RMN

Le lissage des données est indispensable lorsqu'un *binning* a préalablement été effectué puisque cela a pour conséquence de transformer les pics des spectres en signaux plus « triangulaires ». La méthode la plus couramment utilisée pour le lissage des données spectrales est celle de Savitzky-Golay dont le principe est expliqué en Annexe G [148].

L'identification des zones d'intérêt du spectre est également préconisée. En effet, de nombreuses zones spectrales ne font apparaitre que du bruit de fond. Le fait de ne pas en tenir compte dans l'analyse permet de réduire considérablement le nombre de variables. Par exemple, dans le cas des produits issus de l'HCK des DSV, la zone correspondant aux déplacements chimiques compris entre 117 et 150 ppm, caractéristique des carbones aromatiques ne contient en général pas de signaux significatifs, ce type de composé étant très peu présents dans ces produits. C'est le cas du spectre de coupe huile représenté sur la Figure 19. Dans ce cas, la zone d'intérêt se situe entre 0 et 60 ppm (zone caractéristique des carbones aliphatiques saturés), le pic du solvant $CDCl_3$ situé à 76,9 ppm n'étant pas pris en compte puisque tous les échantillons d'une même base de données sont en général analysés avec les mêmes proportions. Ce choix a pour effet de réduire la taille du vecteur d'intensités de 65536 à 13926 valeurs.

La normalisation des spectres est souvent préconisée pour donner à chacun d'eux un poids équivalent dans l'analyse multivariée. Cette étape permet de plus de compenser des variations entre les spectres dues à l'instabilité de l'instrument ou à la préparation de l'échantillon [15,73]. Pour notre étude nous avons choisi d'appliquer à nos bases de données spectrales le prétraitement suivant :

1. Normalisation de chaque spectre en divisant les signaux initiaux par la somme totale des intensités du spectre considéré ;
2. Limitation de la zone spectrale à [0 ;60] ppm ;
3. *Binning* des spectres par méthode de Sousa *et al.* [143] ;
4. Lissage par méthode de Savitzky-Golay.

L'évolution des spectres ^{13}C RMN des 31 échantillons de coupe huile durant les étapes du prétraitement dans la zone comprise entre 29 et 31 ppm est illustrée successivement sur la Figure 21, la Figure 22 et la Figure 23. On observe une nette amélioration du décalage entre les spectres initiaux (Figure 21) et les spectres après *binning* et lissage (Figure 23) qui souligne l'efficacité des méthodes employées.

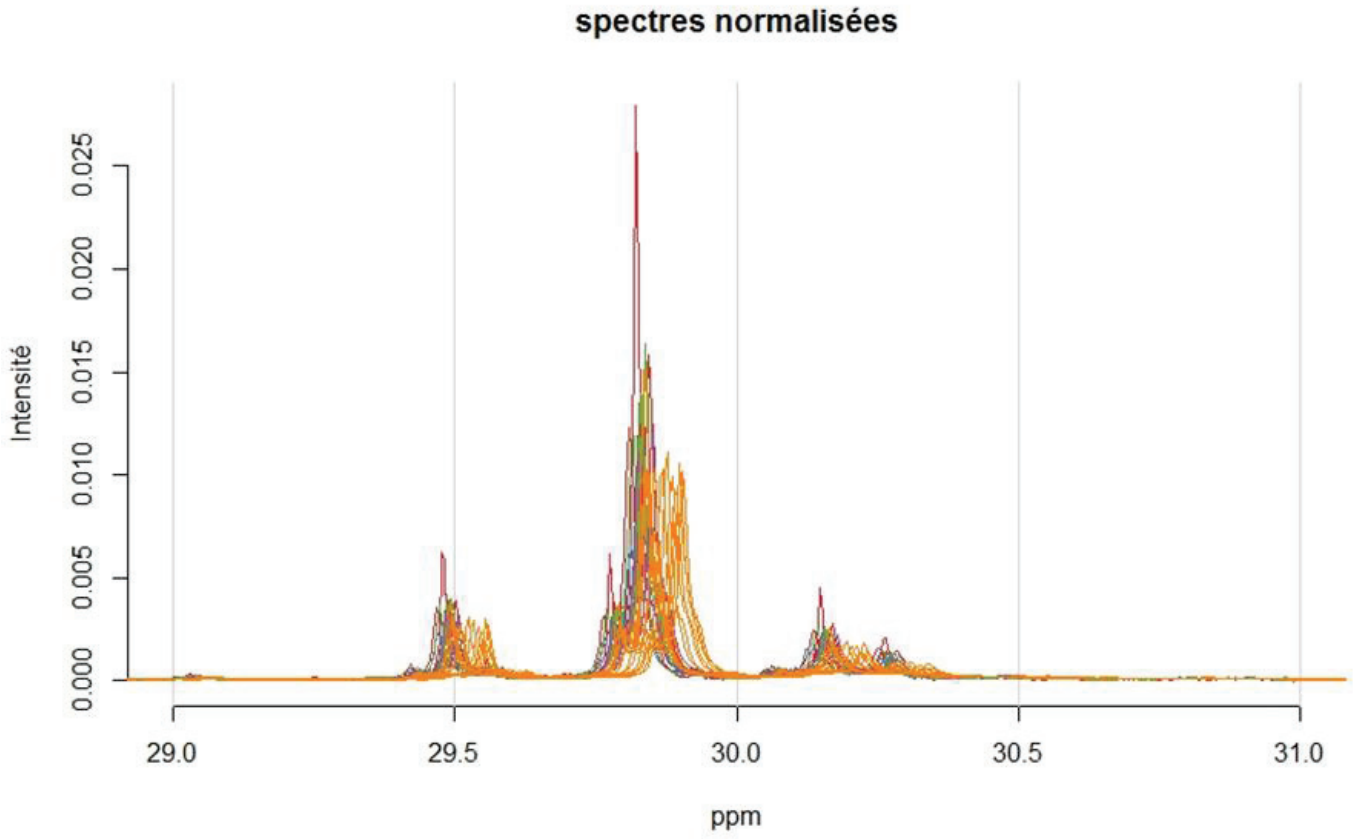


Figure 21 : Illustration des étapes de prétraitement des spectres dans la zone [29 ;31] ppm - superposition des spectres normalisées

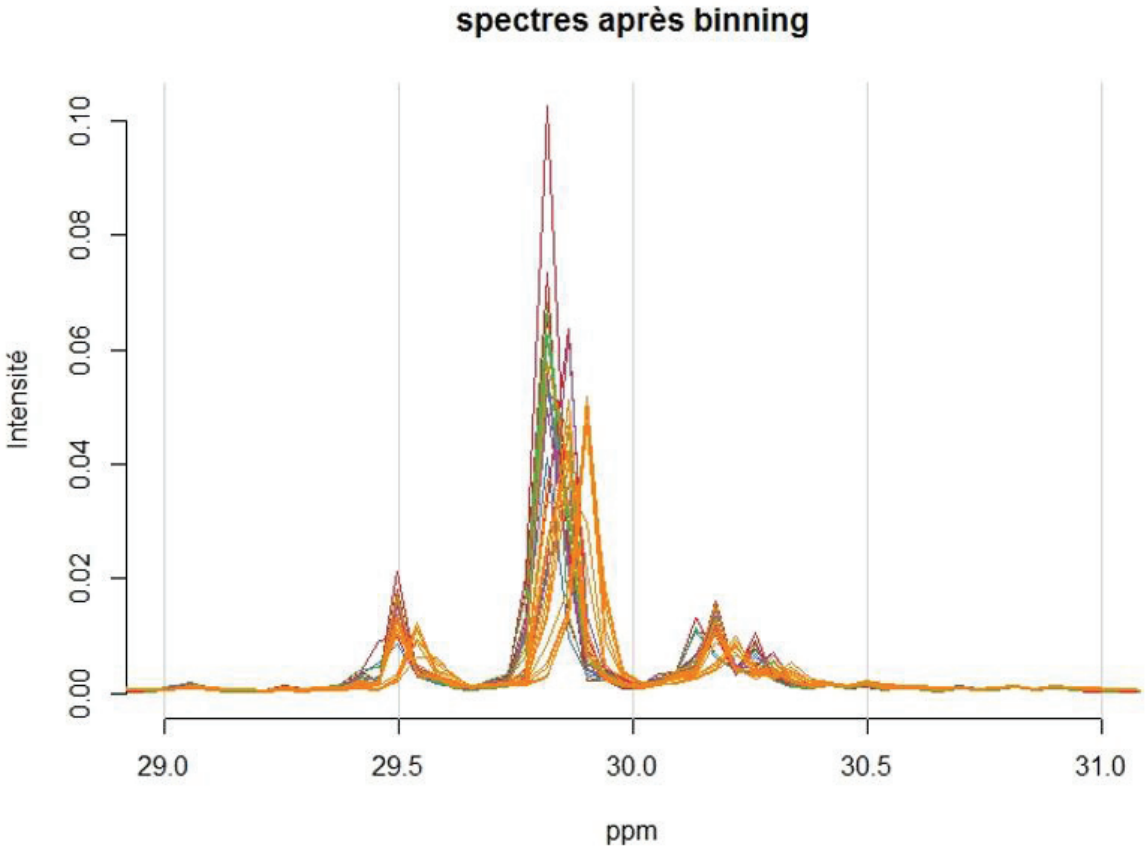


Figure 22 : Illustration des étapes de prétraitement des spectres dans la zone [29 ;31] ppm - superposition des spectres normalisées après *binning* ;

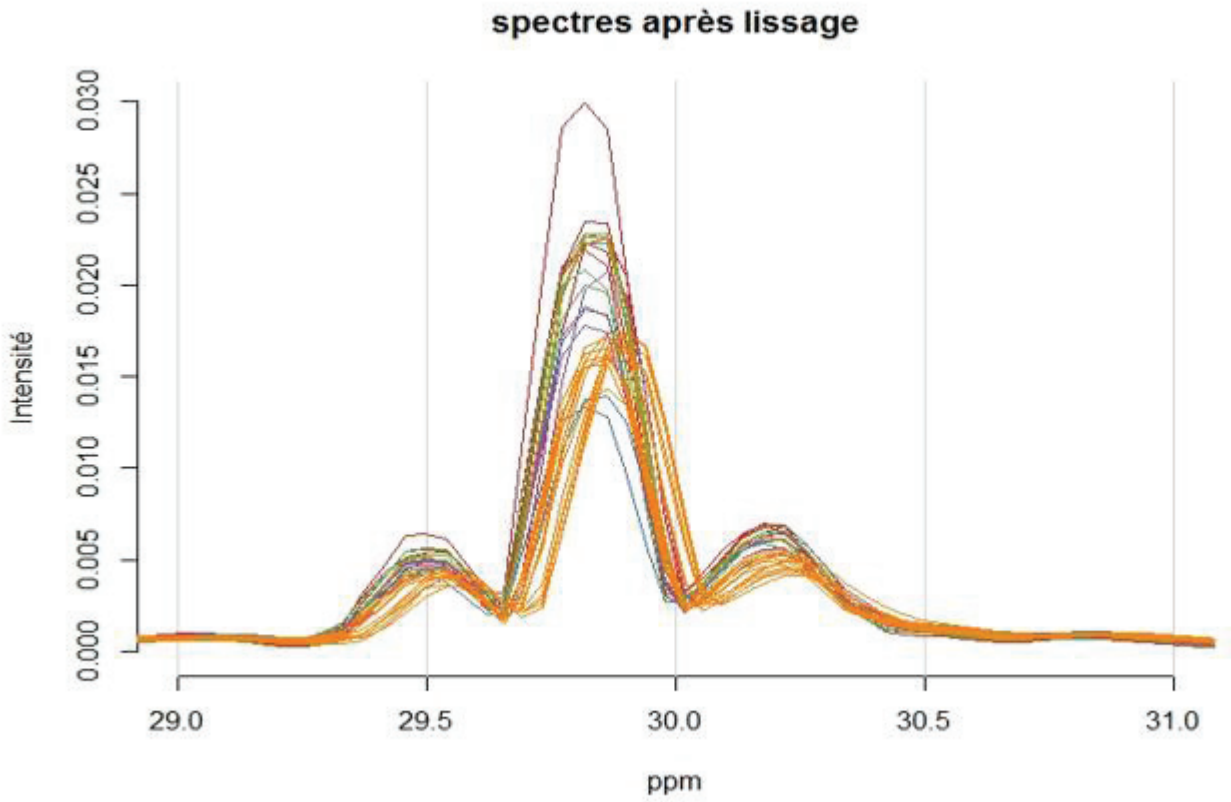


Figure 23 : Illustration des étapes de prétraitement des spectres dans la zone [29 ;31] ppm - superposition des spectres normalisées après *binning* et lissage

Chapitre 4. Méthodes statistiques et chimiométriques

Dans le chapitre précédent, nous avons rappelé les principes de méthodes de caractérisation de coupes pétrolières et le processus de mise en forme des données à exploiter. Le présent chapitre revient sur les aspects théoriques des méthodes statistiques et chimiométriques qui ont été utilisées pour la modélisation du PT de la coupe gazole et VI de la coupe huile. Nous présenterons tout d'abord les méthodes d'analyse exploratoire de données multivariées qui ont été mise en œuvre pour décrire les données. Ensuite, nous reviendrons sur les méthodes prédictives qui ont été développées pour modéliser les propriétés d'intérêt. Les principaux fondements de la RLM et du modèle gaussien seront rappelés, de même que ceux des régressions PLS et *sparse* PLS [52,53]. Nous introduirons de plus les aspects théoriques liés aux méthodes d'interpolation à travers le krigeage et la méthode des *splines*. Une comparaison de ces deux approches d'interpolation sur des jeux de données simulées est également proposée.

4.1 Méthodes d'analyse exploratoire de données

Les principaux objectifs de l'analyse exploratoire de données (multidimensionnelles ou multivariées) sont de résumer ces données, les représenter graphiquement, réduire la dimensionnalité et les regrouper suivant des critères. Les méthodes de statistique exploratoire multidimensionnelle se décompose en deux grands groupes selon l'objectif fixé :

1. Les méthodes dites factorielles de décomposition qui consistent à projeter les données sur une base adaptée [149–151],
2. Les méthodes algorithmiques qui visent la recherche de classes, ou regroupements d'individus les plus proches au sens d'une mesure de distance [152].

Dans ce paragraphe, nous décrivons succinctement les méthodes d'analyse exploratoire de données quantitatives qui ont été utilisées au cours de nos travaux.

4.1.1 Analyse en composantes principales

L'idée principale de l'analyse en composantes principales (ACP) est de réduire l'espace d'étude d'un ensemble de données contenant un grand nombre de variables inter-corrélées, tout en retenant un maximum de variabilité de ces données [149,153]. Pour obtenir ce résultat on substitue aux variables initiales de nouvelles variables appelées « composantes principales » (PCs) qui sont décorrélées entre elles (contraintes d'orthogonalité), et qui sont ordonnées de sorte que la première composante contient la plus grande part de la variance de l'ensemble des données initiales.

4.1.2 Classification ascendante hiérarchique

La classification ascendant hiérarchique (CAH) est une méthode qui consiste à regrouper itérativement les individus, en commençant par les deux plus proches et en construisant progressivement

un arbre ou dendrogramme, regroupant finalement tous les individus en une seule classe [154,155]. L'arbre ou dendrogramme est construit progressivement de sorte que les niveaux de distance augmentent de bas en haut. Chaque regroupement nécessite de savoir calculer la distance entre deux individus ou entre un individu et un groupe, ou encore entre deux groupes (les groupes ayant été formés lors des étapes précédentes). Dans cette étude nous utiliserons uniquement la distance euclidienne dont les règles sont les suivantes [154] :

1. La distance entre un individu et un groupe est le minimum des distances entre l'individu et chaque individu appartenant au groupe
2. La distance entre deux groupes est la plus petite distance entre tous les couples d'individus entre ces deux groupes.

Un exemple de CAH est illustré sur la Figure 24 pour un jeu de données simulées bidimensionnelles. Les deux points les plus proches (1 et 5) sont regroupés lors de la première étape notée (a). L'étape (b) regroupe les points 4 et 6 (la distance entre 6 et 4 est la plus faible après l'étape (a)). Les autres étapes regroupent successivement :

- le point 3 et le groupe formé par 4 et 6 (étape (c))
- le groupe formé par 1 et 5 et celui formé par 3, 4 et 6 (étape (d))
- le point 2 et groupe formé par 1, 3, 4, 5, 6 (étape (e)).

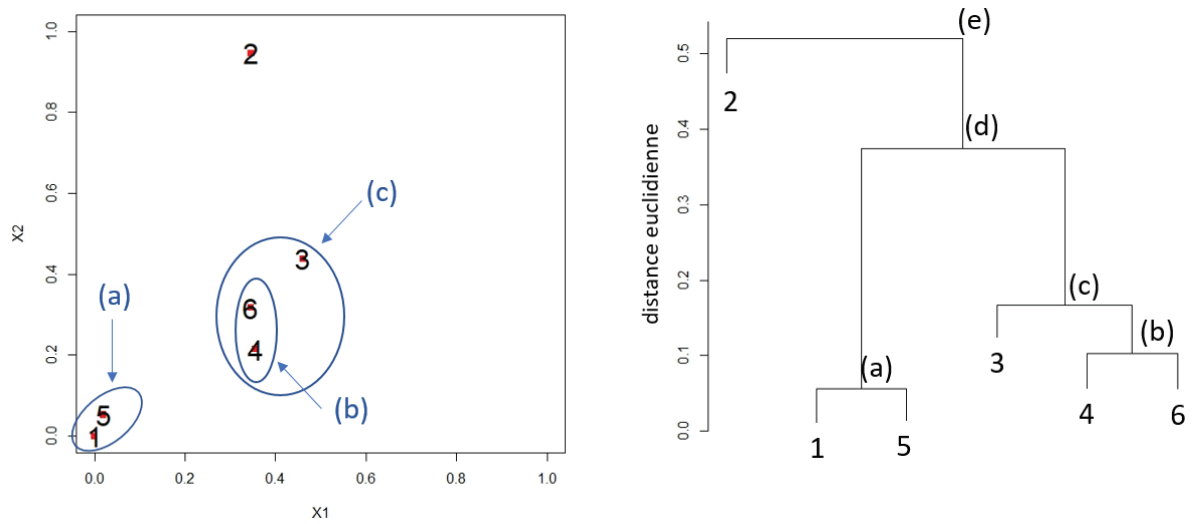


Figure 24 : Exemple de CAH sur un jeu de données simulées ; (à gauche) Répartition des points dans le plan (X1, X2) ; (à droite) dendrogramme correspondant (distance euclidienne)

4.2 Méthodes de sélection de variables

La sélection de variables descriptives est une étape essentielle dans le développement de modèles prédictifs. En modélisation, on est souvent confronté à la recherche d'un équilibre entre la qualité de l'ajustement (le biais) et la variance des paramètres afin d'assurer à la fois la justesse et la

précision et de garantir de bonnes capacités prédictives au modèle (éviter le surapprentissage). Dans le cas des modèles dits « gaussiens » telles que la RLM, de nombreuses méthodes ont été développées qui permettent de trouver un bon compromis entre ces deux phénomènes. Elles peuvent être classées en trois groupes :

1. Les méthodes algorithmiques (*backward, forward, stepwise*) [42–45]
2. Les méthodes d'optimisation de critères pénalisés (C_p, *AIC, BIC,* R^2 *ajusté*) [46,47,156]
3. Les méthodes parcimonieuses (ridge, Lasso, *elastic net*) [49,55,157]

Les méthodes algorithmiques et les méthodes parcimonieuses sont surtout utilisées dans le cas où le nombre de descripteurs potentiels est très grand (données spectrales ou chromatographiques par exemple) où elles sont plus efficaces car elles ne requièrent pas de tester un grand nombre de combinaisons.

Des méthodes alternatives plus adaptées à des modèles complexes ont été introduites. C'est le cas de l'analyse de sensibilité qui repose sur le calcul d'indices permettant de quantifier l'influence d'une variable d'entrée d'un modèle donné sur la sortie (paragraphe 4.2.2) [51,158,159].

Pour notre étude, deux méthodes de sélection de variables ont été mises en oeuvre :

- Sélection par maximisation du R^2 *ajusté*
- Sélection par analyse de sensibilité

4.2.1 Sélection de variables par optimisation de critères

Comme leur nom l'indique, les méthodes d'optimisation de critères consiste à choisir la combinaison de descripteurs optimale, c'est-à-dire celle qui minimise ou maximise un critère prédéfini. Dans notre étude, nous avons utilisé le critère du R^2 *ajusté*. Cette statistique est une version pénalisée du coefficient de corrélation classique noté R^2. Elle est définie comme suit [46] :

$$\boldsymbol{R^2_{ajusté} = 1 - \frac{(1-R^2)(n-1)}{n-1-k}} \qquad \textbf{(Eq. 2. 5)}$$

où n désigne la taille de la base d'apprentissage et k le nombre de variables sélectionnés. Dans le cas du R^2 *ajusté* il s'agit d'un problème de maximisation. Le principal défaut du R^2 est qu'il croît avec le nombre de descripteurs. Or R^2 *ajusté* est à la fois une fonction croissante de R^2 et une fonction décroissante de k. Il faut donc trouver la combinaison de variable qui correspond à un k pas trop grand et un R^2 pas trop petit. D'où l'intérêt de la méthode.

4.2.2 Sélection de variable par analyse de sensibilité

L'analyse de sensibilité consiste à quantifier l'impact d'un descripteur donné sur la variable à prédire [51,158,159]. Elle permet de classer les différents descripteurs suivant leur influence sur la sortie du modèle.

Supposons que l'on dispose d'un système complexe modélisé par une fonction f qui décrit la relation entre l'ensemble des variables explicatives x et la réponse y, *i.e.* :

$$\boldsymbol{f(x) = y} \qquad \textbf{(Eq. 2. 6)}$$

Pour un descripteur x_i donné, on définit la sensibilité comme une mesure de la variabilité de y lorsque x_i est égal à une certaine valeur. On définit alors pour un descripteur donné l'indice de sensibilité du premier ordre noté S_i comme suit [51] :

$$\boldsymbol{S_i = \frac{Var(\mathbb{E}[(y|x_i)])}{Var(y)} \in [0; 1]} \qquad \textbf{(Eq. 2. 7)}$$

où $\mathbb{E}[(y|x_i)]$ désigne la valeur moyenne prise par y lorsqu'on fixe la variable x_i. Des indices de sensibilité d'ordre supérieur ou égal à 2 ont également été définis [51]. Pour notre étude, nous nous sommes limités à l'ordre 1 car nos modèles ne possèdent pas de termes d'interaction entre les variables. **La méthode de sélection de variables par analyse de sensibilité consiste donc à sélectionner celles qui ont les indices de sensibilité d'ordre 1 les plus hauts.** Une interprétation théorique des S_i est proposée en Annexe H.

4.3 Régression linéaire multiple

Dans ce paragraphe nous proposons une approche de la régression linéaire multiple (RLM) basée sur l'ouvrage de Hastie *et al.* [57] et sur celui d'Azaïs et Bardet [46].

4.3.1 Approche théorique de la RLM

Pour fixer les idées, nous noterons par la suite :

- y la variable à modéliser (propriété, champ, *etc.*)
- $x = (x_1, \dots, x_p)$ un ensemble de p variables explicatives (conditions expérimentales, données analytiques, *etc.*)
- x^T le vecteur transposé de x

Nous supposerons de plus que nous disposons de n mesures de y, notées $(y_1, \dots, y_n)$ prise pour différentes valeurs de x. Enfin, nous noterons X la matrice de données expérimentales, c'est-à-dire la matrice qui contient les échantillons en lignes et les variables en colonnes, un élément $x_{i,j}$ de la matrice contient la valeur mesurée pour la $j^{ème}$ variable du $i^{ème}$ échantillon. On parle de modèle RLM lorsque la fonction de modélisation s'écrit sous la forme :

$$\boldsymbol{f(x, \theta) = \theta_0 + \sum_{i=1}^{p} \theta_i x_i} \qquad \textbf{(Eq. 2. 8)}$$

où f est la fonction de modélisation et les θ_i sont les paramètres du modèle à estimer. Pour les modèles de régression, l'étape d'apprentissage consiste à estimer « au mieux » l'ensemble des paramètres θ à partir des mesures expérimentales dont on dispose. Par la suite, nous noterons $\hat{\theta}$ tout estimateur des paramètres θ du modèle (voir notion d'estimateur en Annexe D)**.** L'estimation des paramètres d'un modèle RLM abouti en général à la résolution d'un problème d'optimisation dont la fonction coût notée $J(\theta)$ dépend des observations. La forme de cette fonction détermine la nature de l'estimateur (moindres carrés, maximum de vraisemblance, Ridge, Lasso, *etc.*). L'estimateur des moindres carrés est le plus utilisé. Il consiste à minimiser la somme des carrés des résidus du modèle. Sous les hypothèses d'un modèle gaussien, l'estimateur des moindres carrés présente des caractéristiques d'optimalité (il est sans biais et de variance minimale [46]). Le modèle est dit gaussien si l'erreur de modélisation est prise comme une variable aléatoire de distribution normale d'espérance nulle et de variance inconnue notée σ^2. On montre de plus que l'estimateur des moindres carrés peut s'écrire sous la forme :

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{LS}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y_1}, \dots, \boldsymbol{y_n})^{\boldsymbol{T}} \qquad \textbf{(Eq. 2. 9)}$$

4.3.2 Incertitude prédiction pour les modèles RLM gaussiens

L'approche usuelle de modélisation des incertitudes de prédiction en un point x_0 donné est basée sur les hypothèses de distribution gaussienne des erreurs de modélisation. Les intervalles de confiance à 95% associés à la valeur prédite sont alors définis comme suit :

$$\boldsymbol{Prob}\{\boldsymbol{y}(\boldsymbol{x_0}) \in [\widehat{\boldsymbol{y}}(\boldsymbol{x_0}) - \boldsymbol{2}\,\boldsymbol{\sigma}(\boldsymbol{x_0}), \widehat{\boldsymbol{y}}(\boldsymbol{x_0}) + \boldsymbol{2}\,\boldsymbol{\sigma}(\boldsymbol{x_0})]\,\} = \boldsymbol{0,95} \qquad \textbf{(Eq. 2. 10)}$$

où $\sigma(x_0)$ désigne l'écart type de l'incertitude de prédiction. Dans le cas de modèle RLM, on montre qu'on peut estimer cet écart-type par :

$$\widehat{\boldsymbol{\sigma}}^{\boldsymbol{2}}(\boldsymbol{x_0}) = \widehat{\boldsymbol{o}}^{\boldsymbol{2}}(\boldsymbol{1} + \boldsymbol{x_0}^{\mathbf{T}}\left(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x_0}) \qquad \textbf{(Eq. 2. 11)}$$

où $\hat{\sigma}^2$ est un estimateur sans biais de la variance de l'erreur de prédiction. Il est défini comme suit [46] :

$$\widehat{\boldsymbol{o}}^{\boldsymbol{2}} = \frac{\boldsymbol{1}}{\boldsymbol{n-1}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_{\boldsymbol{i}})^{\boldsymbol{2}} \qquad \textbf{(Eq. 2. 12)}$$

Ces intervalles de confiance dits « gaussien » ont une amplitude qui croît avec la distance entre x_0 et le centre de gravité de la base d'apprentissage. Par ailleurs, cette croissane est relativement lente, ce qui conduit en première approximation à considérer ces intervalles comme étant d'amplitude homogène.

4.4 Régression PLS et *sparse PLS*

La *sparse* PLS est une variante de la régression PLS classique qui est une méthode de régression multivariée très utilisée en chimiométrie [52]. Dans ce paragraphe, nous présentons les fondements théoriques de ces méthodes en pointant la différence entre les deux approches.

4.4.1 Régression PLS

La régression PLS est l'approche la plus utilisée de la PLS en chimiométrie [53,160]. Le principe de la méthode est de substituer aux variables explicatives initiales (généralement des données spectrales ou chromatographiques) un nombre $r < n$ de facteurs de substitution, communément appelés « variables latentes ». Ces dernières sont construites de sorte qu'elles soient liées autant que possible à la variable à prédire (au sens de la covariance empirique). Tout comme les composantes principales, les variables latentes sont orthogonales entre elles et sont des combinaisons linéaires des variables initiales. Considérons la matrice d'expérience X (de taille $n \times p$ avec $n < p$). On cherche une matrice U de coefficients (appelées *loadings*) qui définisse les r variables latentes ($\Gamma_h, h = 1, \dots, r$) par des combinaisons linéaires des colonnes de X, *i.e.* :

$$\boldsymbol{\Gamma = XU} \qquad \textbf{(Eq. 2. 13)}$$

On montre que la matrice U est solution du problème suivant :

$$\begin{cases} \boldsymbol{Pour\ h = 1, \dots, r,\ \ u_h = \arg\max_u Cov(y, \Gamma_h)^2} \\ \boldsymbol{sous\ contraintes :\ u_h^T u_h = 1\ et\ \Gamma_h^T \Gamma_l = 0\ si\ l \neq h} \end{cases} \qquad \textbf{(Eq. 2. 14)}$$

où u_h et Γ_h désignent respectivement les $h^{ème}$ colonnes des matrices U et Γ. La matrice des *loadings* est obtenue par démarche itérative. Il suffit ensuite de calculer la régression de y sur les r variables latentes Γ_h ainsi construites. Le nombre r est généralement optimisé par validation croisée. Plusieurs algorithmes itératifs ont été définis pour déterminer la matrice U. Dans notre cas, l'algorithme utilisé est issu des travaux de Wold *et al.* [53].

Le fait que les variables latentes soient des combinaisons linéaires des variables initiales constitue l'inconvénient principal de la méthode. En effet, l'interprétation physique des variables latentes s'avère le plus souvent complexe.

4.4.2 Principe de la *sparse* PLS

La *sparse* PLS permet de surmonter la difficulté liée à l'interprétabilité des modèles en introduisant une contrainte de parcimonie lors de la construction des variables latentes pour réduire leur complexité [54,161]. La parcimonie est une notion très utilisée dans la résolution de problème numérique en très grande dimension. Elle est basée sur l'hypothèse que lorsqu'un système est décrit par un grand nombre de variables, très peu d'entre elles influent en réalité sur une propriété donnée. Cette contrainte se traduit mathématiquement par l'introduction d'une pénalisation de type Lasso [49] dans l'algorithme de construction des variables latentes. La matrice U des *loadings* est maintenant déduite par :

$$\begin{cases} \boldsymbol{Pour\ h = 1, \dots, r, u_h = \arg\max_u Cov(y, \Gamma_h)^2 + \eta \|u_h\|_1} \\ \dots \end{cases} \qquad \textbf{(Eq. 2. 15)}$$

où $\|.\|_1$ désigne une norme vectorielle définie comme la somme des valeurs absolues des coefficients d'un vecteur donné et η est appelé degré de parcimonie. Ce type de pénalisation conduit à l'annulation

des coefficients les plus petits pour ne laisser qu'un ensemble restreint de coefficients non nuls dont le nombre dépend de la valeur de η.

Pour la *sparse* PLS, l'étape de validation a donc deux objectifs :

1. Choisir le nombre de variables latentes nécessaire pour obtenir une bonne prédiction
2. Fixer au mieux le degré de parcimonie qui permet de réduire la complexité des variables latentes en gardant l'information essentielle.

En pratique ces deux paramètres sont optimisés simultanément par validation croisée. Dans notre cas, le degré de parcimonie sera optimisé pour un nombre de variables latentes équivalent à celui du modèle de régression PLS.

4.5 Méthodes d'interpolation

Le principe des méthodes d'interpolation est d'estimer une propriété en un point par une moyenne pondérée de valeurs mesurées aux voisinages de ce point. Elles offrent donc plus de flexibilité que les méthodes de régression classiques. Il existe deux types de méthodes d'interpolation [58] : les méthodes déterministes (interpolation par *splines*) qui ont pour but de créer des surfaces à partir de points mesurés, en fonction du degré de similarité (pondération par l'inverse de la distance) ou du degré de lissage (fonctions de base radiale) ; et les méthodes stochastiques (krigeage) qui quantifient l'autocorrélation spatiale entre les points de mesures et tiennent compte de la configuration spatiale des points d'échantillonnage autour de l'emplacement de prévision. Tout comme les méthodes d'apprentissage non paramétrique, les méthodes d'interpolation ne requièrent pas de définir au préalable une forme analytique de la fonction de modélisation. Elles présentent par ailleurs l'avantage d'être plutôt efficace dans les cas où on dispose d'un nombre relativement réduit d'observation.

4.5.1 Interpolation par krigeage simple

Le krigeage est une méthode d'interpolation probabiliste couramment utilisée en géostatistique. L'approche proposée dans ce paragraphe est inspirée de l'ouvrage de Pierre Goovaerts [62].

4.5.1.1 Illustration sur un cas tridimensionnel

Pour illustrer le principe de la méthode on se place dans le cas où la réponse à modéliser y dépend de deux variables explicatives (x_1, x_2). On suppose qu'on dispose d'une dizaine d'observation de y correspondants à des couples (x_1, x_2) distincts. La Figure 25 illustre une situation d'interpolation dans le cas décrit ci-dessus. Les axes horizontaux et verticaux désignent les domaines des variables explicatives (x_1, x_2) et la réponse y est représentée par une échelle de couleur (bleu pour les valeurs les plus faibles et rouges pour les plus fortes). Les points d'observations correspondants à la base d'apprentissage sont en couleur (elles ont été numérotées de 1 à 10) et le point à prédire est en blanc. Le krigeage consiste à estimer y en ce point comme une moyenne pondérée d'observations situées dans un voisinage proche, délimité sur la Figure 25 par l'ellipse en vert. Ainsi, sur cet exemple, seules les observations 4, 5 et 6 seront utilisées pour l'estimation de y en (?). L'attribution des poids se fait de plus

en fonction de la distance entre les sites d'observation et le point à prédire. Plus cette distance est grande, plus faible sera le poids [62].

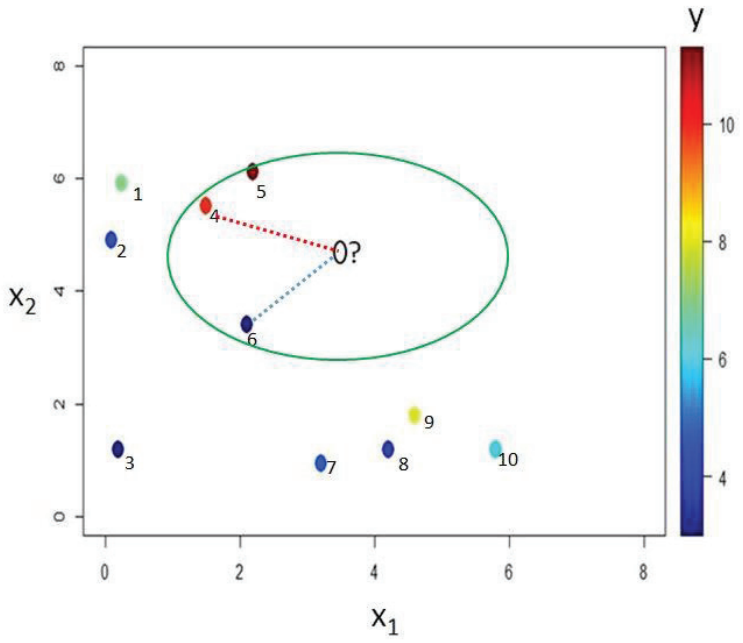


Figure 25 : Illustration d'un problème d'interpolation dans un cas tridimensionnel

4.5.1.2 Extension au cas multidimensionnel

Notons s_0 le point où l'on veut prédire y et $\{s_1, \dots, s_n\}$ l'ensemble des points de la base d'apprentissage. Les s_i sont donc des $p-uplet$ qui représentent chacun une suite de coordonnées. Notons enfin $\hat{y}^K(s_0)$ l'estimateur par krigeage de y en s_0. Alors il peut s'écrire sous la forme :

$$\hat{y}^K(s_0) = \sum_{j \in J(s_0)} p_j y(s_j) \qquad \textbf{(Eq. 2. 16)}$$

où $J(s_0)$ est le sous-ensemble de $\{1, \dots, n\}$ qui contient les indices des observations qui sont situées à l'intérieur du voisinage d'influence de s_0. Les poids p_j associés aux observations retenues dépendent de trois paramètres :

1. La configuration géométrique des données d'apprentissage
2. La distance entre les sites d'observations et le nouveau point
3. Les caractéristiques structurales de la réponse (isotropie, régularité, *etc.*)

En pratique, la distribution des poids est régie par un modèle mathématique appelé « modèle de covariance » entre les données d'apprentissage. Notez que le krigeage est *a priori* une méthode exacte, c'est-à-dire que l'erreur de modélisation est nulle en tout point de la base d'apprentissage.

4.5.1.3 Approche statistique du krigeage

L'approche statistique sur laquelle repose les fondements du krigeage est différente de celle des modèles de régression et notamment du modèle RLM (paragraphe 4.3). Elle est basée sur une

approche stochastique qui consiste à modéliser la réponse y en tout point s de l'espace d'étude comme la somme d'une fonction déterministe notée μ et d'une fonction aléatoire δ, *i.e.* :

$$\boldsymbol{y(s) = \mu(s) + \delta(s)} \qquad \textbf{(Eq. 2. 17)}$$

L'hypothèse fondamentale du krigeage est que la fonction aléatoire δ décrit un processus stochastique multigaussien sur l'ensemble de l'espace d'étude, c'est-à-dire que pour tout site s, $\delta(s)$ est une variable aléatoire gaussienne et que l'ensemble des variables aléatoires ainsi définies sont *a priori* corrélées entre elles [162]. On suppose de plus que ce processus est stationnaire d'ordre 2 : Ces deux hypothèses implique que d'une part la valeur moyenne de $\delta(s)$ est la même en tout point s de l'espace d'étude et d'autre part que la covariance entre deux variables aléatoires prises en deux sites distincts ne dépend que de l'écart entre ces sites, *i.e.* pour tout site s et pour tout réel h :

$$\begin{cases} \boldsymbol{E[\delta(s_i)] = m} \\ \boldsymbol{Cov(\delta(s), \delta(s+h)) = C(h)} \end{cases} \qquad \textbf{(Eq. 2. 18)}$$

où C est une fonction mathématique qui définit le modèle de covariance entre les sites. Suivant la forme de μ il existe trois types de krigeage :

- Le krigeage simple (SK) où μ est une fonction constante de valeur connue
- Le krigeage ordinaire (OK) où μ est une fonction constante de valeur inconnue
- Le krigeage universel (UK) où μ peut s'écrire comme une combinaison linéaire de fonctions polynomiales, *i.e.* il existe une suite de fonctions $(f_j)_{j=1,\dots,p}$ telles que

$$\boldsymbol{\mu(s) = \sum_{j=1}^{p} f_j(s)\alpha_j} \qquad \textbf{(Eq. 2. 19)}$$

En général le krigeage simple et le krigeage ordinaire sont équivalents. Dans notre étude, seul le krigeage simple a été utilisé.

4.5.1.4 Prédiction par krigeage simple

Dans ce cas on suppose que y s'écrit sous la forme :

$$\boldsymbol{y(s) = m + \delta(s)} \qquad \textbf{(Eq. 2. 20)}$$

où m est une constante réelle. De ce fait, **toutes les hypothèses faites sur $\boldsymbol{\delta(s)}$ sont également valables pour $\boldsymbol{y(s)}$.** La prédiction par krigeage simple consiste à rechercher l'estimateur linéaire optimal (sans biais et de variance minimale,

) de $y(s_0)$ sous les hypothèses définies au paragraphe précédent. Notons $\hat{y}^{SK}(s_0)$ l'estimateur par krigeage simple de $y(s_0)$. Ce dernier doit donc vérifier les trois contraintes suivantes :

$$\begin{cases} \boldsymbol{linéarité}: \ \boldsymbol{\hat{y}^{SK}(s_0) = \sum_{i=1}^{n} p_i y(s_i)} \\ \boldsymbol{sans\ biais}: \ \boldsymbol{E\left[\hat{y}^{SK}(s_0)\right] = y(s_0)} \\ \boldsymbol{variance\ minimale}: \ \boldsymbol{\hat{y}^{SK}(s_0) = \min_{\hat{y}^L(s_0)} Var[\hat{y}^L(s_0)]} \end{cases} \qquad \textbf{(Eq. 2. 21)}$$

où $\boldsymbol{\hat{y}^L(s_0)}$ désigne tout estimateur linéaire de $y(s_0)$. Les p_i sont appelés « poids du krigeage ». Sous les hypothèses de stationnarité (paragraphe 4.5.1.3), et en combinant les équations, on montre que pour tout point x_0 de l'espace d'étude, les poids du krigeage p_i sont solutions du système linéaire suivant :

$$\begin{bmatrix} \boldsymbol{C(0)} & \cdots & \boldsymbol{C(s_1 - s_n)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{C(s_n - s_1)} & \cdots & \boldsymbol{C(0)} \end{bmatrix} \begin{bmatrix} \boldsymbol{p_1} \\ \vdots \\ \boldsymbol{p_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C(s_1 - s_0)} \\ \vdots \\ \boldsymbol{C(s_n - s_0)} \end{bmatrix} \qquad \textbf{(Eq. 2. 22)}$$

Notez que le système ci-dessus et donc les poids du krigeage dépendent à la fois du point où l'on veut prédire la propriété et de la structure de covariance entre les sites d'observations. Il existe plusieurs formes de fonctions de covariance (gaussienne, exponentielle, linéaire, sphérique, *etc*.). Les expressions analytiques des modèles les plus courants sont précisées dans le Tableau 11. Dans le cas du krigeage, l'étape analogue à la phase d'apprentissage consiste à estimer au mieux les paramètres intrinsèques de la structure de covariance. Le critère d'optimisation le plus couramment utilisé dans cette optique est la maximisation de la fonction de vraisemblance [163,164].

Tableau 11 : Exemples de structure de covariance disponibles dans la littérature [165]

Structure de covariance	Expression
Exponentielle	$C(s_i, s_k) = \sigma_0^2 e^{-\theta_j (x_{j,i} - x_{j,k})}$
Gaussienne	$C(s_i, s_k) = \sigma_0^2 e^{-\sum_{j=1}^{p} \theta_j (x_{j,i} - x_{j,k})^2}$
Linéaire	$C(s_i, s_k) = max\{0{,}1 - \theta_j \lvert x_{j,i} - x_{j,k} \rvert\}$
Sphérique	$C(x_i, x_k) = 1 - 1.5\xi_j + 0.5{\xi_j}^3, \ \xi_j = \min\{1, \theta_j \lvert x_{j,i} - x_{j,k} \rvert\}$

θ_j: paramètres de la structure de covariance ; σ_0^2: variance de y supposée homogène sur l'ensemble de l'espace d'étude ; $x_{j,i}$: $j^{ème}$ coordonnée de s_i dans l'espace d'étude

La résolution du système donné dans l'Eq. 2.18 fournit les poids optimaux et permet d'obtenir $\hat{y}^{SK}(s_0)$. Par ailleurs, cette solution est en adéquation avec les aspects du krigeage qui ont été présentés au paragraphe 4.5.1.2. En effet, supposons que s_0 est un site de la base d'apprentissage, par exemple $s_1 = s_0$. La première ligne du système défini par l'équation Eq. 2.18 devient alors :

$$\boldsymbol{C(0)p_1 + C(s_1 - s_2)p_2 + \cdots + C(s_1 - s_n)p_n = C(0)} \qquad \textbf{(Eq. 2. 23)}$$

Une solution évidente de cette équation est de prendre $p_1 = 1$ et $p_i = 0$ pour tout $i > 1$. De plus, la fonction C est généralement construite de telle sorte que la matrice du système soit symétrique définie

positive et donc inversible, ce qui assure l'unicité de la solution du système pour un site s_0 donné. Ainsi, dans le cas où $s_1 = s_0$ l'estimateur par krigeage s'écrit simple s'écrit :

$$\hat{y}^{SK}(s_0) = \sum_{j=1}^{n} p_j y(s_j) = y(s_1) \qquad \textbf{(Eq. 2. 24)}$$

Cette démonstration peut être étendue à tout autre site d'observation. La fonction krigée passe bien par les mesures des sites d'observation.

4.5.1.5 Incertitude de prédiction pour modèles stochastiques

La plupart des méthodes statistiques prédictives sont déterministes. De ce fait, elles ne fournissent pas d'incertitudes relatives à leurs prévisions. L'approche probabiliste offre des possibilités plus intéressantes. L'approche stochastique est différente de l'approche gaussienne classique. Notons $u(s_0)$ la variable aléatoire qui modélise l'incertitude de prédiction au point s_0. On définit de plus pour tout réel z :

$$F(s_0, z) = Prob\{u(s_0) \leq z | (I)\} \qquad \textbf{(Eq. 2. 25)}$$

comme étant la fonction de répartition de $u(s_0)$ conditionnellement à l'ensemble des informations utilisées pour la construction du modèle (base de données d'apprentissage par exemple) noté(I). L'approche stochastique consiste à construire l'intervalle de confiance de la prédiction *via* cette distribution, *i.e. :*

$$Prob\{u(s_0) \in [a; b] | (I)\} = F(s_0, a) - F(s_0, b) \qquad \textbf{(Eq. 2. 26)}$$

où a et b sont des grandeurs scalaires. Notez que ces intervalles de confiance stochastiques ne dépendent pas de l'estimateur de $y(s_0)$ considéré, mais uniquement des informations contenues dans (I) et du point s_0. On montre par ailleurs que sous les hypothèses de distribution multi-gaussienne liées au krigeage simple (paragraphe 4.5.1.3), **la distribution de $u(s_0)$ ainsi définie est d'espérance nulle et de variance égale à celle de $\hat{y}^{SK}(s_0)$ [62]**.

4.5.2 Interpolation par *splines*

La méthode des *splines* [59,60,166,167] est une méthode d'interpolation déterministe qui consiste à approcher la variable à modéliser par des morceaux de polynômes locaux de degré m. Les *splines* peuvent être linéaires, quadratiques ou cubiques ($m = 1$, 2 ou 3 respectivement). L'approche la plus classique de la méthode consiste à approcher la variable à modéliser par la fonction $\hat{y}^{sp}$ qui est solution du problème :

$$\hat{y}^{sp} = \min_{\phi} \left\{ J_m(\phi) = \sum_{\alpha_1+\cdots+\alpha_p=m} \frac{m!}{\alpha_1!\alpha_2!\ldots\alpha_p!} \int \cdots \int_{\mathbb{R}^p} \left(\frac{\partial^m \phi}{\partial x_1^{\alpha_1} \ldots \partial x_p^{\alpha_p}} \right)^2 dx_1 \ldots dx_p \right\} \qquad \textbf{(Eq. 2. 27)}$$

La fonction $J_m(\phi)$ peut être interprétée en première approximation comme une « énergie de flexion » d'une plaque fine (en dimension $p > 1$). Donc, la fonction ϕ qui minimise cette énergie prend la forme d'une plaque fine qui passe obligatoirement par les points de la base d'apprentissage. En pratique, on

utilise une approche un peu différente appelée « *smoothing spline* ». Elle consiste à substituer le problème de minimisation précédent par :

$$\widehat{\boldsymbol{y}}^{\boldsymbol{sp}} = \underset{\boldsymbol{\phi}}{\boldsymbol{min}} \left\{ \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{w_i^2} \left[\boldsymbol{\phi}\left(\boldsymbol{x_{1,i}}, \ldots, \boldsymbol{x_{p,i}}\right) - \boldsymbol{y_i} \right]^{\boldsymbol{2}} + \boldsymbol{\lambda J_m}(\boldsymbol{\phi}) \right\} \qquad \textbf{(Eq. 2. 28)}$$

où λ est appelé paramètre de lissage et les poids w_i^2 sont fixés par l'utilisateur. Le terme rajouté a pour conséquence de forcer la solution du problème à ne pas passer trop loin des points de la base de calibration tandis que le paramètre de lissage permet de jouer sur la flexibilité de la solution du problème. Pour la méthode des *splines*, l'étape de validation consiste donc à optimiser à la fois le paramètre λ et l'ordre m des polynômes locaux. La forme de la solution du problème associé à l'interpolation par *splines* est discutée dans [60].

4.6 Outils statistiques pour l'évaluation des performances de modèle

Dans ce paragraphe nous définissons les outils statistiques qui seront utilisés pour évaluer les performances des modèles de prédiction. Ces statistiques sont récapitulées dans le Tableau 12. y désigne toujours la réponse à modéliser. Notez que nous avons volontairement introduit une distinction entre les termes « réel » et « mesuré » (Tableau 12). $y^{réel}$ désigne la vraie valeur de la réponse en un point donné tandis que $y^{mesuré}$ désigne la valeur observée ou mesurée. Ces deux valeurs sont rigoureusement différentes puisqu'en pratique les mesures sont généralement accompagnées d'incertitudes. La valeur réelle de la réponse n'est vraiment connue que dans le cas de données simulées. Ainsi, dans le cas de la modélisation de propriétés physicochimiques, les valeurs de RMSE et de MAD, calculées par rapport à la valeur mesurée seront toujours associées au pourcentage de point prédits avec une erreur inférieure (en valeur absolue) à l'incertitude de la mesure (Tableau 12). La valeur de γ sera fixée en fonction des fidélités de la méthode de mesure standard pour une propriété donnée. Dans le cas où les valeurs de la variable à prédire sont relativement faibles, il peut être judicieux de raisonner en termes d'erreur relative moyenne (MRD). Toutes ces statistiques sont estimées aussi bien sur la base d'apprentissage que sur la base de test.

Tableau 12 : Statistiques pour l'évaluation des performances des modèles

Statistiques	Formule
Erreur absolue moyenne	$MAD = \frac{1}{n}\sum_{i=1}^{n}\left\|y_i^{prédit} - y_i^{réel\ ou\ mesuré}\right\|$
Erreur relative moyenne	$MRD = \frac{1}{n}\sum_{i=1}^{n}\frac{\left\|y_i^{prédit} - y_i^{réel\ ou\ mesuré}\right\|}{y_i^{réel\ ou\ mesrué}}$
Erreur quadratique moyenne	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{prédit} - y_i^{réel\ ou\ mesuré})^2}$
Pourcentage de points prédits avec une erreur absolue inférieure à γ	$\tau_{\pm\gamma} = \frac{\sum_{i=1}^{n} 1_{\left\|y_i^{prédit} - y_i^{réel\ ou\ mesuré}\right\| \leq \gamma}}{n}$

n : nombre de points évalués

4.7 Comparaison de modèles : régression vs interpolation

L'objet de ce paragraphe est d'effectuer une comparaison entre les modèles de RLM et les méthodes d'interpolation par *splines* et par krigeage. L'objectif de cette comparaison est d'une part d'illustrer les avantages de l'utilisation des méthodes d'interpolation par rapport aux méthodes de régression, et d'autre part de mettre en avant l'utilité de l'approche probabiliste qu'offre le krigeage. Ces points sont discutés à travers trois exemples de données simulées.

4.7.1 Bases de données simulées

Les trois bases de données simulées sont présentées ci-dessous. Dans chaque cas, on se placera dans le cas de données tridimensionnelles. Les bases de données ont été simulées suivant le processus ci-dessous :

1. Définition de l'espace d'étude D partie de $\mathbb{R}^2$,
2. Génération aléatoire d'un ensemble de coordonnées bidimensionnelles dans D
3. Définition de l'expression analytique de la fonction f à modéliser
4. Application de f à chaque point généré dans l'étape 2 et ajout d'un bruit blanc (variable aléatoire de loi normale centrée et de variance σ^2)

Ainsi, chaque base de données est constituée d'un ensemble de n triplet (x, y, z) tels que :

$$\boldsymbol{z = f(x,y) + \varepsilon, (x,y) \in D \subset \mathbb{R}^2 \ et\ \varepsilon \sim N(0, \sigma^2)} \qquad \textbf{(Eq. 2. 29)}$$

Le Tableau 13 donne pour chaque cas l'expression analytique de la fonction à modéliser, le domaine d'étude et la variance du bruit. Trois fonctions ont été choisies pour cette étude : une fonction affine, une fonction rationnelle et une fonction dite de potentiel.

Tableau 13 : Expressions analytiques des fonctions à modéliser

Type de fonction	Expression analytique	D	σ^2
Affine	$z = f(x,y) = 2x - 3y + 4$	$[-3;3] \times [-3;3]$	1
Rationnelle	$z = f(x,y) = \dfrac{(x^3 - 3x)}{1 + y^2}$	$[-3;3] \times [-3;3]$	1
Fonction de potentiel	$z = f(x,y) = \dfrac{0{,}3}{0{,}3 + (x+0{,}5)^2 + (y+0{,}5)^2} + \dfrac{0{,}7}{0{,}7 + (x-0{,}5)^2 + (y-0{,}5)^2}$	$[-1;1] \times [-1;1]$	0,025

4.7.1.1 Fonction affine (Jeu 1)

Dans le premier cas, la fonction à modéliser est affine (Tableau 13). La base de données est constituée de 50 points uniformément répartis sur l'ensemble de l'espace d'étude $[-3;3]^2$ et de la valeur bruitée de f en chacun de ces points. L'ensemble des données simulées est représenté sur la Figure 26.

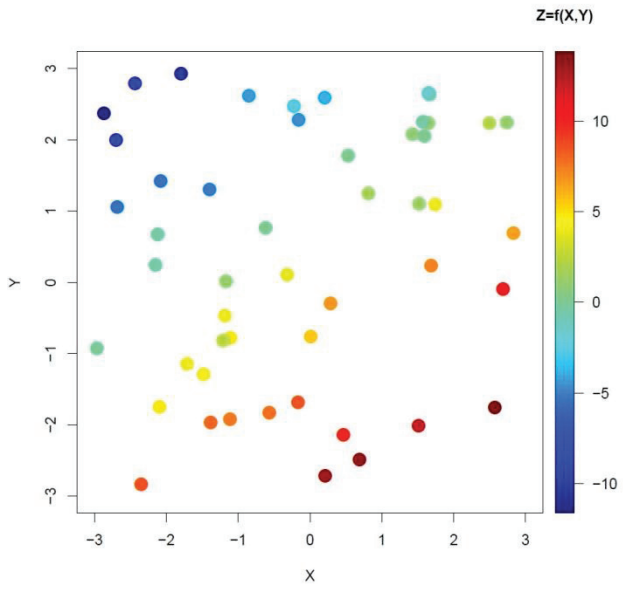


Figure 26 : Représentation de la base de données simulée dans le cas de la fonction affine

4.7.1.2 Fonction rationnelle (Jeu 2)

Dans ce second cas, la fonction à modéliser est une fraction rationnelle (quotient de polynôme) dont l'expression analytique est donnée dans le Tableau 13. La base de données est constituée de 50 points

toujours uniformément répartis sur $[-3;3]^2$. La répartition des données sur le plan (x,y) est donnée sur la Figure 27.

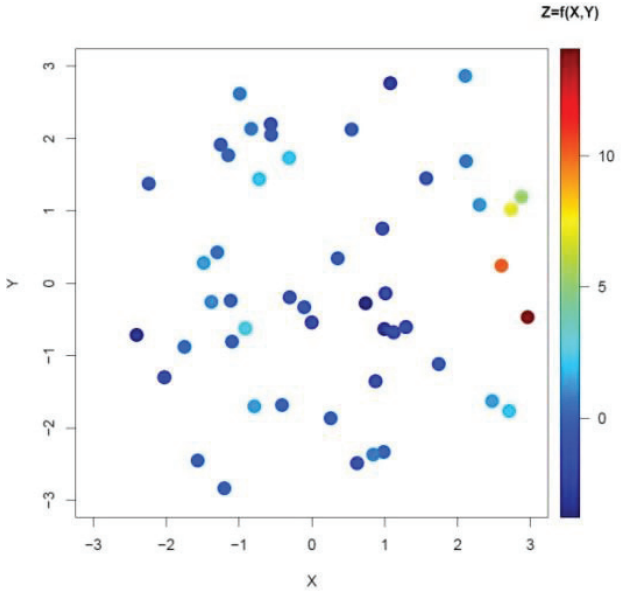


Figure 27 : Représentation de la base de données simulée dans le cas de la fonction rationnelle

4.7.1.3 Fonction de potentiel (Jeu 3)

Dans ce troisième cas, la fonction à modéliser peut être interprétée comme la somme de deux potentiels centrés en deux points P_1 et P_2 symétriques par rapport à l'origine du repère de l'espace d'étude et de coordonnées respectives $(-0{,}5;-0{,}5)$ et $(0{,}5;0{,}5)$. Les données ont été générées de sorte qu'elles forment deux clusters bien identifiés. Pour cela, chaque couple de coordonnées de l'espace d'étude $[-1;1]^2$ a été simulé suivant une loi normale bidimensionnelle centrée soit en P_1 :

$$(x_i, y_i) \sim \mathcal{N}\left(\begin{pmatrix}-0.5\\-0.5\end{pmatrix}; \begin{pmatrix}0.25 & 0\\0 & 0.25\end{pmatrix}\right) \qquad \text{(Eq. 2. 30)}$$

soit en P_2 :

$$(x_i, y_i) \sim \mathcal{N}\left(\begin{pmatrix}-0.5\\-0.5\end{pmatrix}; \begin{pmatrix}0.25 & 0\\0 & 0.25\end{pmatrix}\right) \qquad \text{(Eq. 2. 31)}$$

La base de données ainsi simulée est représentée sur la Figure 28.

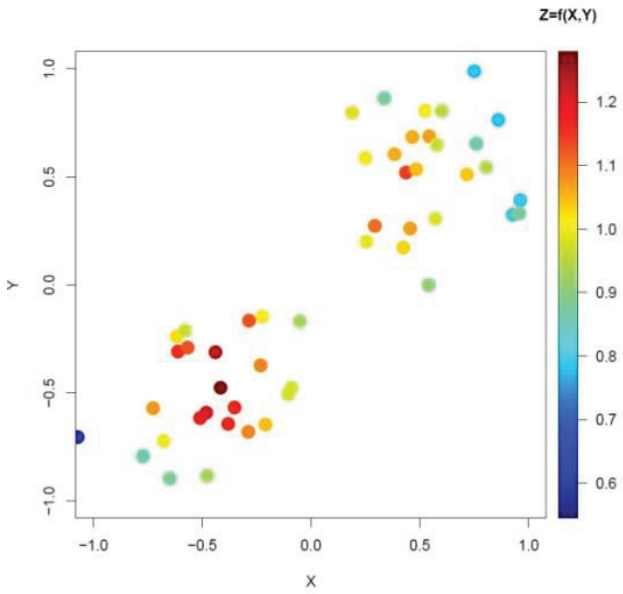


Figure 28 : Représentation de la base de données simulée dans le cas de la fonction de potentiel

4.7.2 Evaluation des performances

Pour chaque jeu de données présenté au paragraphe précédent, l'espace d'étude a été discrétisé en une grille de 10000 points (100×100) régulièrement répartis. Ces points ont été utilisés comme base de test pour évaluer les performances des modèles développés. Des statistiques présentées au paragraphe 4.6 ont ainsi été calculées.

4.7.3 Résultats et discussions

Pour chaque fonction simulée au paragraphe précédent trois modèles ont été développés *via* la RLM (paragraphe 4.3), la méthode des *splines* (paragraphe 4.5.2) et le krigeage (paragraphe 4.5.1). Les données simulées ont été utilisées comme base d'apprentissage pour la calibration des modèles. Les différents résultats obtenus sont présentés et discutés ci-dessous.

4.7.3.1 Modélisation de la fonction affine (Jeu 1)

Pour le jeu de données 1 (paragraphe 4.7.1.1) les résultats sont illustrés sur la Figure 29. Sur la Figure 29a, la fonction à modéliser (en noir) et la surface de réponse (en bleu) obtenue à partir de la RLM ont été superposées. Les données d'apprentissage sont représentées par des points verts. On peut noter que les deux surfaces (planes) sont très proches l'une de l'autre. La Figure 29b est une représentation 2D de la répartition des erreurs de modélisation sur l'espace d'étude. La valeur absolue des erreurs de modélisation est indiquée par une échelle de couleur (en rouge les valeurs les plus hautes et en bleu les plus faibles). Les données de calibration sont cette fois représentées par des petits cercles blancs. On observe clairement que l'ensemble de l'espace d'étude est coloré en bleu ce qui signifie que les erreurs de prédiction sont faibles (<0,75). Ces deux figures montrent que la qualité de la prédiction est très bonne, ce qui confirme les propriétés d'optimalité de l'estimateur des moindres carrés dans le

cas du modèle Gaussien (*i.e.* cet estimateur est sans biais et de variance minimale). Toutefois, la surface de réponse fournie par la RLM est rigoureusement distincte du graphe initial. En effet, on peut observer un contraste bleu/noir sur le graphe de superposition des deux courbes (Figure 29a) qui traduit une très légère inclinaison de la surface de réponse par rapport au graphe original de la fonction modélisée.

Une représentation similaire a été faite pour illustrer les résultats obtenus par krigeage et par méthode de *splines*. Dans le cas du krigeage, la superposition de la surface de réponse et de la fonction à modéliser est illustrée sur la Figure 29c. Là encore on observe une forte proximité entre les deux courbes. Ce constat est appuyé par le graphe 2D des erreurs absolues (Figure 29d) puisque la majeure partie de l'espace d'étude est colorée bleu. Cependant, des différences peuvent être notées par rapport à la RLM. Premièrement, la ligne de contraste bleu/noir, non linéaire qu'on observe sur la Figure 29c illustre que la surface krigée est légèrement incurvée. Cela explique la moins bonne qualité de la prédiction du modèle de krigeage par rapport au modèle RLM. Les résultats obtenus sur les *splines* sont similaires à ceux obtenus sur le krigeage (Figure 29e et Figure 29f).

Globalement, les trois méthodes de modélisation sont performantes. Ce constat est appuyé par les statistiques qui sont données dans le Tableau 14. En effet, des MRD de 0,07, 0,13 et 0,10 ont été respectivement obtenues pour la RLM, le krigeage et la méthode des *splines*.

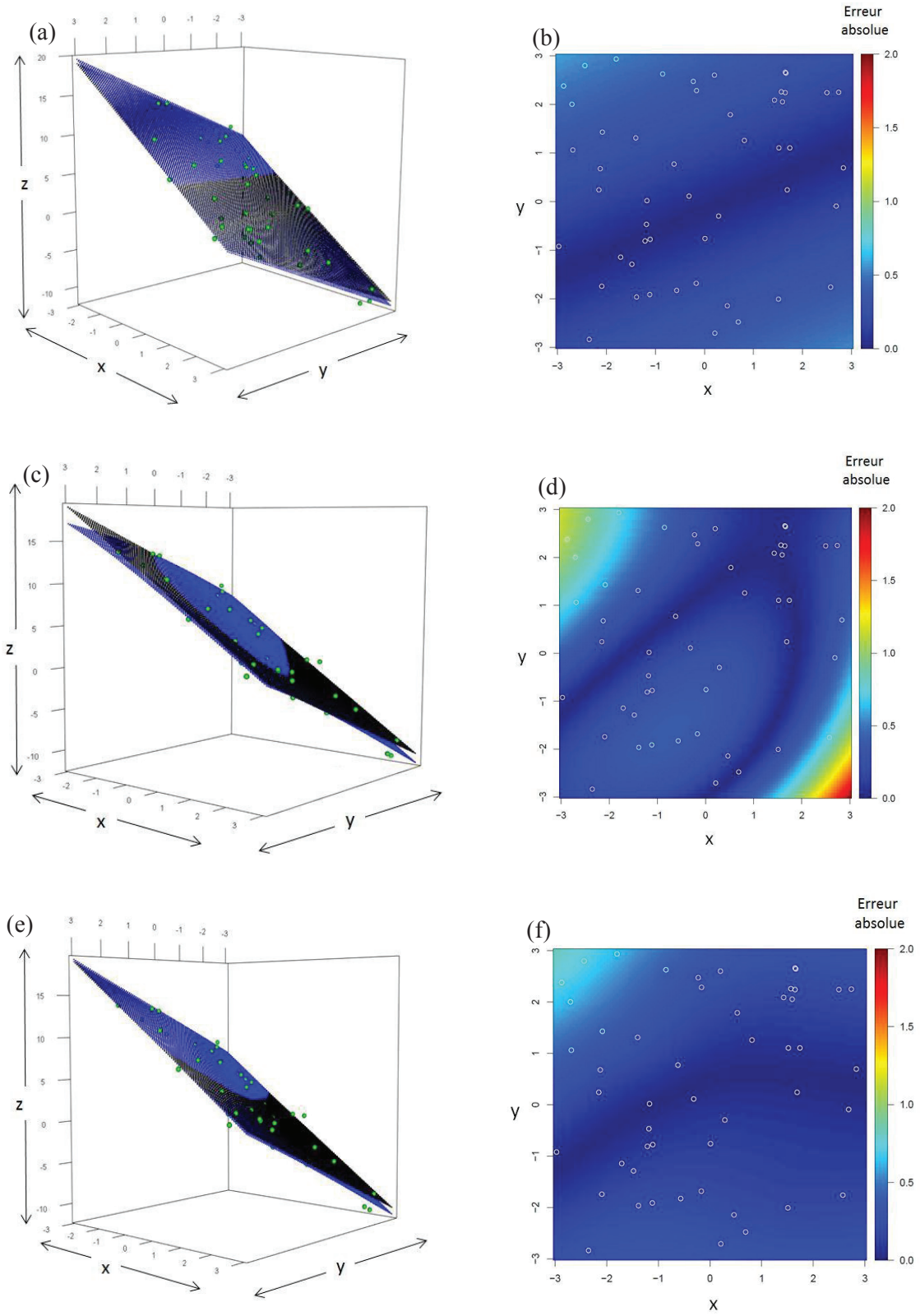


Figure 29 : Superposition de la surface de réponse obtenue par RLM (a), krigeage (c) et *splines* (e) (bleue) et de la fonction affine (noire) ; b) Erreur absolue en tout point de l'espace d'étude pour le modèle RLM (b), de krigeage (d), de *splines* (f) ;

Tableau 14 : Statistiques globales des modèles de prédiction pour chaque jeu de données

Jeu de données	Modèle	MRD	RMSEP
Jeu 1 (fonction affine)	RLM	0,07	0,24
	Krigeage	0,13	0,40
	Splines	0,10	0,23
Jeu 2 (fonction rationnelle)	RLM	7,80	2,52
	Krigeage	3,83	1,88
	Splines	3,60	1,69
Jeu 3 (fonction de potentiel)	RLM	0,40	0,32
	Krigeage	0,10	0,18
	Splines	0,22	0,29

4.7.3.1 Modélisation de la fonction rationnelle (Jeu 2)

Dans ce cas, la fonction à modéliser est une fraction rationnelle. En tenant compte de la forme de ce type de fonction, il est évident que la RLM ne permet pas d'obtenir de solution satisfaisante et cette technique ne sera donc pas appliquée.

Les résultats obtenus à partir de l'interpolation par *splines* et par krigeage sont illustrés la Figure 30. La Figure 30a montre la superposition de la surface krigée et du graphe original de la fonction à modéliser et la Figure 30b montre les erreurs absolues de prédiction correspondantes. Le krigeage semble fournir une bonne précision sur la majeure partie de l'espace d'étude. On note que les régions bleues-foncées (Figure 30b) qui correspondent aux points prédits avec une erreur quasi-nulle sont localisées autour des données de calibration (cercles blancs). Par contre, la qualité de la prédiction est particulièrement dégradée dans une région bien localisée caractérisée à la fois par une quasi-absence des données de calibration et une forte variabilité de la fonction à modéliser. Ces remarques peuvent être étendues à la prédiction par *spline*s (Figure 30c et Figure 30d). Les deux modèles sont clairement très proches l'un de l'autre comme le confirment les statistiques (Tableau 14) (les RMSEP sont respectivement de 1,88 pour le krigeage et 1,69 pour les *splines*). Les valeurs élevées d'erreurs relatives moyennes (>3,60) sont dues à la mauvaise qualité de la prédiction sur certains points de l'espace d'étude.

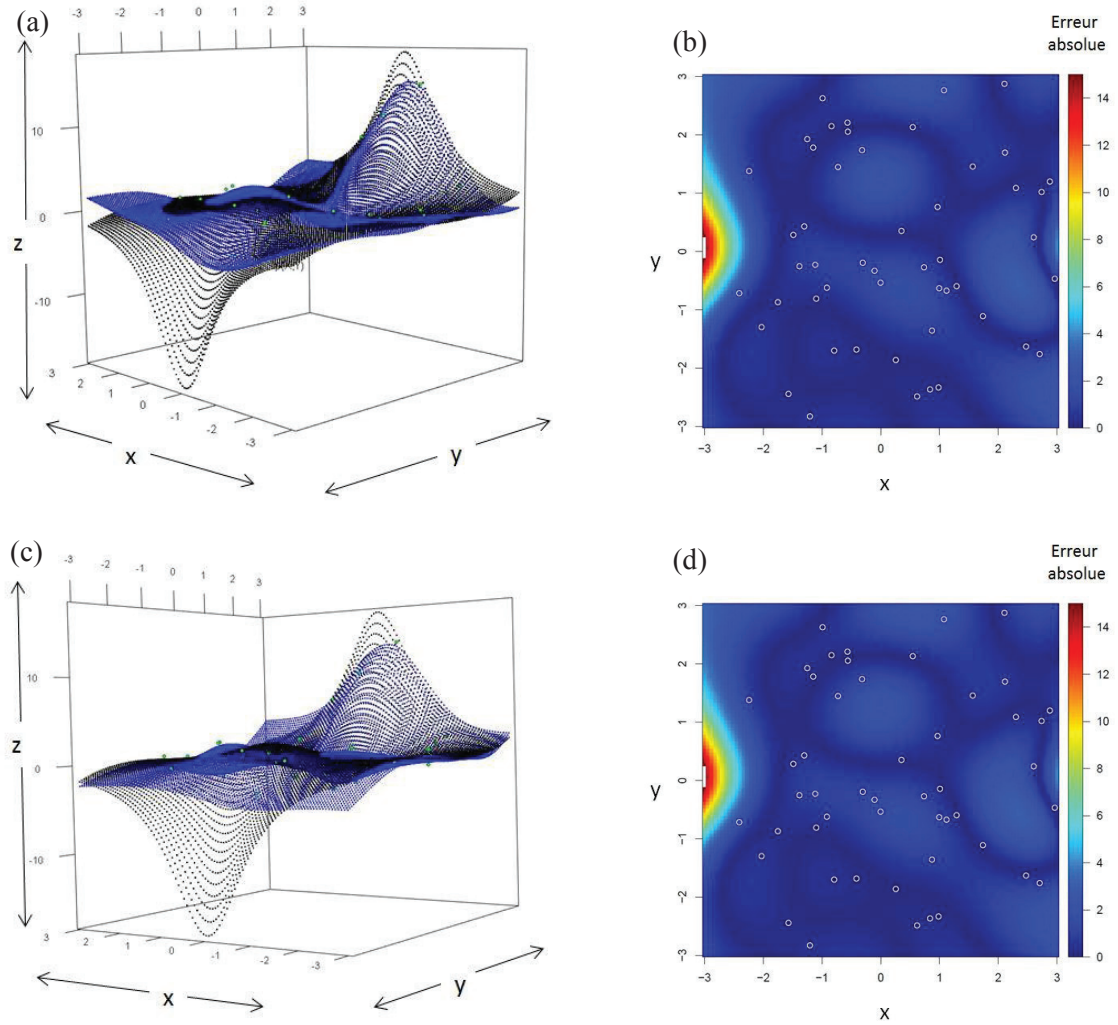


Figure 30 : Superposition de la surface de réponse obtenue par krigeage (a), *splines* (c) (bleue) et de la fonction rationnelle (noire) ; Erreur absolue en tout point de l'espace d'étude pour le modèle de krigeage (b), de *splines* (d) ;

4.7.3.2 Modélisation de la fonction de potentiel (Jeu 3)

Les performances du modèle de krigeage et de *splines* pour ce jeu de données sont illustrées sur la Figure 31. La Figure 31a et la Figure 31c montrent que dans le cas du krigeage comme dans le cas des *splines*, la surface de réponse est très proche du graphe original aussi bien dans les zones à forte affluence de points de calibration que dans la zone située entre les clusters. Ce constat est également visible sur les graphes des résidus absolus (Figure 31b et Figure 31d).

Des différences significatives de comportement des surfaces de réponse sont observables dans les zones où il y a une faible affluence de points. Dans le cas du krigeage, on note une tendance du modèle à surestimer la valeur de la fonction et la surface de réponse semble tendre vers une valeur moyenne (Figure 31a). Dans le cas des *splines* on remarque une sous-estimation et une forte divergence de la surface de réponse par rapport au graphe original (Figure 31c). Le krigeage est ainsi plus

performant dans ce cas (une erreur relative moyenne de prédiction de 0,10 ayant été obtenue contre 0,22 pour la méthode des *splines*).

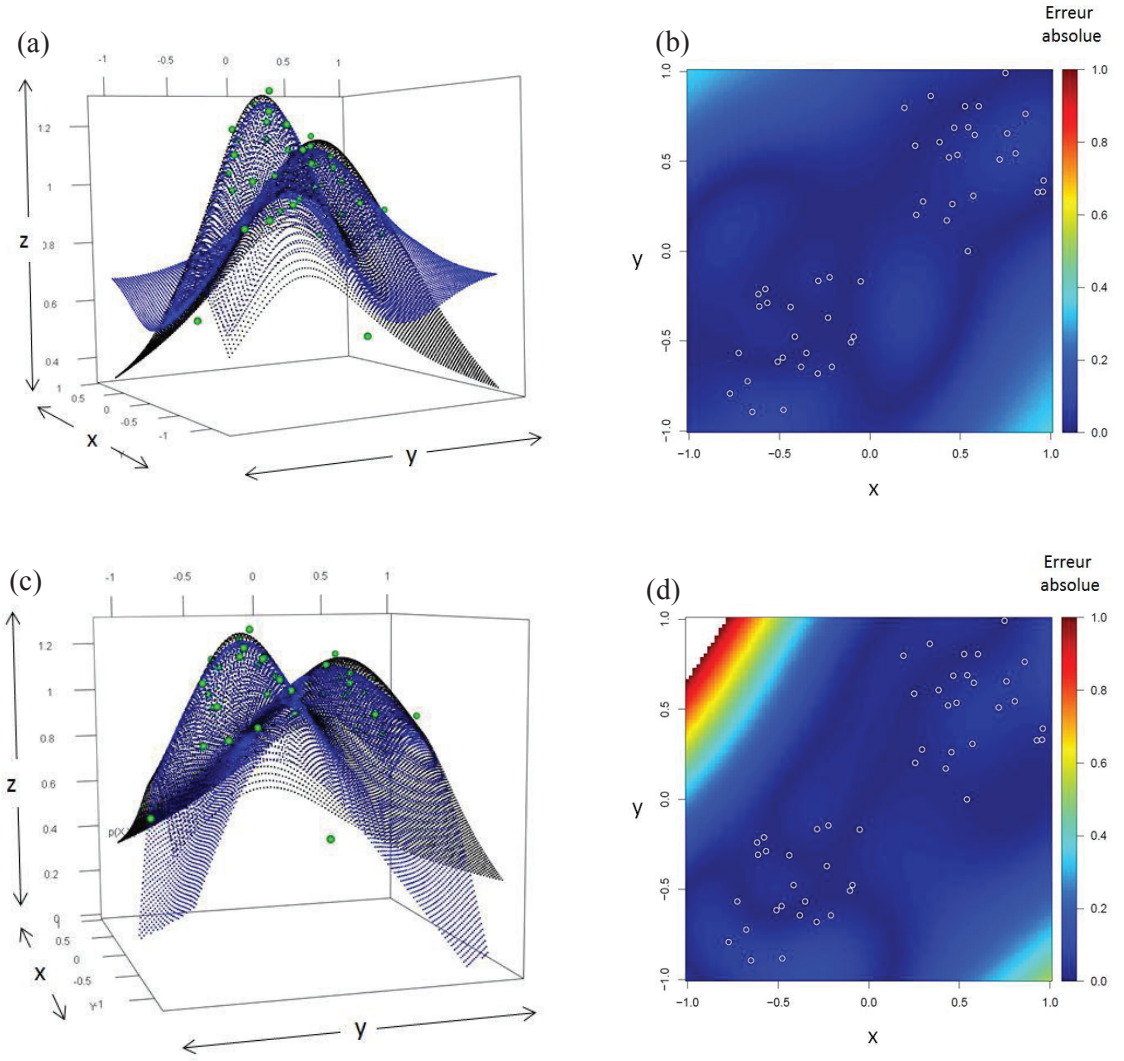


Figure 31 : Superposition de la surface de réponse obtenue par krigeage (a), *splines* (c) (bleue) et de la fonction de potentiel (noire) ; Erreur absolue en tout point de l'espace d'étude pour le modèle de krigeage (b), de *splines* (d) ;

4.7.4 Incertitudes de prédiction

Les intervalles de confiance de prédiction ont été calculées pour chaque jeu de données suivant deux approches : l'approche dite classique qui consiste à assimiler la réponse à une variable gaussienne et qui est couramment utilisée en régression, et l'approche stochastique reliée à la théorie du krigeage (paragraphe 4.5.1.5). Les amplitudes des intervalles de confiance à 95% associés à chaque point de l'espace d'étude sont illustrées sur la Figure 32 dans le cas de la fonction rationnelle et de la fonction de potentiel. Les amplitudes de ces intervalles sont indiquées par une échelle de couleur. Les zones colorées en bleu correspondent aux intervalles de faibles amplitudes (donc de faibles incertitudes de prédiction) et les zones en rouge aux incertitudes les plus grandes. Les cercles blancs représentent toujours les points

de la base d'apprentissage. On note que dans les 2 situations les intervalles de confiance classiques ont la forme d'ellipses dont le centre se confond avec le centre de gravité du nuage de points (Figure 32a et Figure 32c). L'incertitude augmente de manière régulière avec la distance à ce centre de gravité. Les intervalles de prédiction obtenus par l'approche stochastique sont clairement différents. Les régions de plus faibles incertitudes sont clairement localisées autour des points de la base d'apprentissage (Figure 32b et Figure 32d). L'incertitude de prédiction associée à un point dépend de sa position par rapport à ces données.

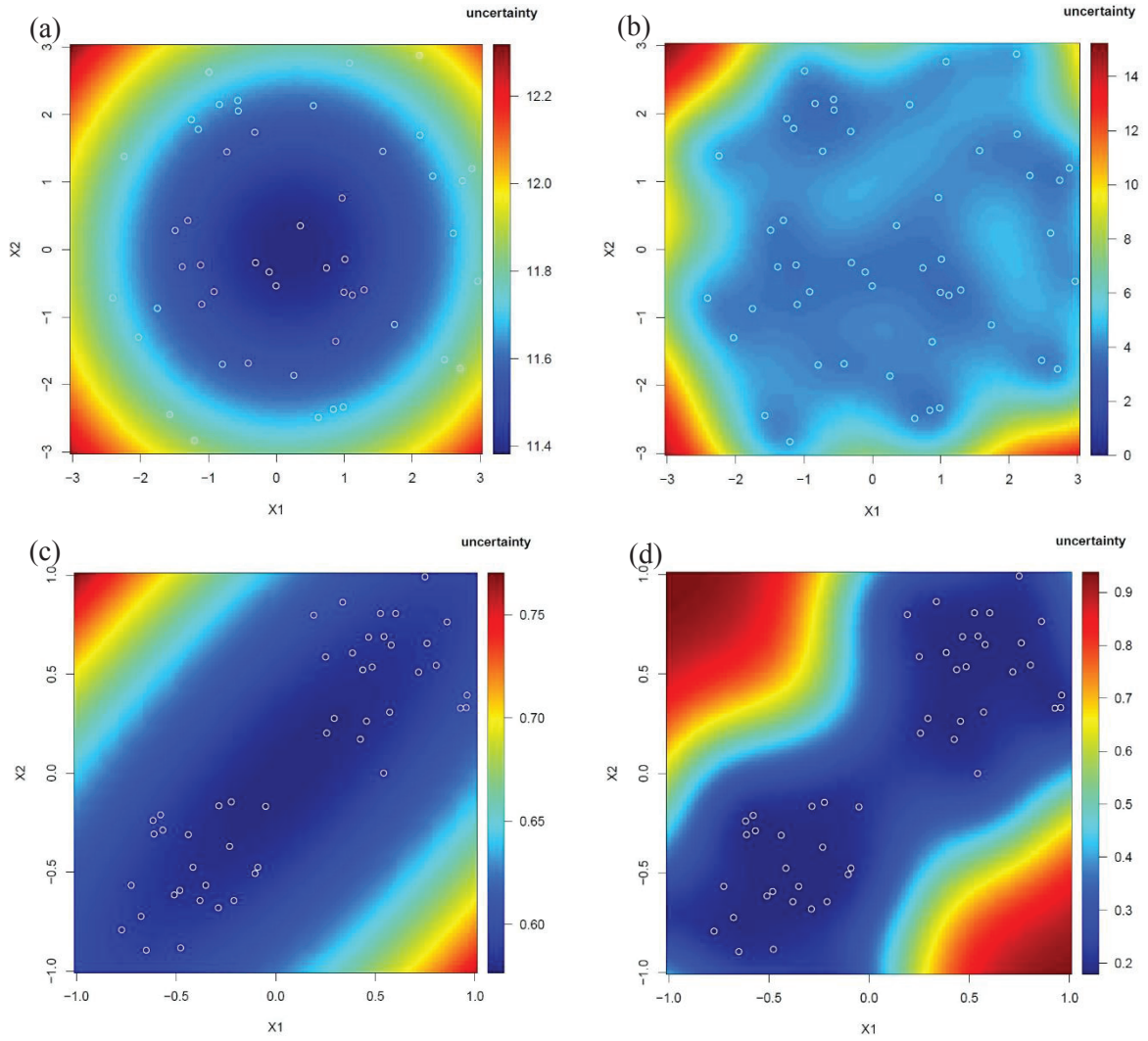


Figure 32 : Amplitudes des intervalles de confiance de prédiction pour la fonction rationnelle par méthode classique (a), méthode stochastique (b) ; amplitudes des intervalles de confiance de prédiction pour la fonction de potentiel par méthode classique (c), méthode stochastique (d)

4.7.5 Bilan

Des modèles de RLM, d'interpolation par krigeage et par *splines* ont été testés sur trois cas de données simulées. En résumé, deux situations distinctes ont été observés. La première concerne les situations d'« interpolation », c'est-à-dire lorsque la prédiction est faite en un point entouré de points

voisins de la base de d'apprentissage. Dans ces cas, le krigeage et la méthode des *splines* donnent naturellement des prédictions très précises. La seconde fait référence aux situations d'extrapolation où le point à prédire se trouve relativement éloigné des données d'apprentissage. On note alors différents comportements des méthodes suivant la fonction à modéliser. Lorsque cette dernière varie faiblement dans la zone localisée autour du point considéré, la précision de la prédiction reste relativement bonne. Par contre, lorsque la fonction à modéliser varie fortement, la prédiction s'écarte de la valeur réelle plus ou moins fortement suivant la méthode considérée. Cette différence de comportement entre les deux méthodes d'interpolation s'explique par le choix de la structure de covariance gaussienne pour le modèle de krigeage. En effet, Dubrule [166] a montré que le krigeage et la méthode des *splines* sont équivalents dans le cas où une structure de type polynomiale (de même ordre que celui des polynômes de *splines*) est utilisée. Cependant, notre étude a montré que ce type de structure n'est pas toujours approprié, notamment dans le cas de la fonction de potentiel où la surface de réponse obtenue par méthode de *splines* diverge fortement du graphe initial (Figure 31c). L'utilisation de la structure gaussienne permet de contenir cette divergence en faisant tendre la surface de réponse vers une valeur moyenne.

Plus généralement, les résultats ont montré la capacité des méthodes d'interpolation à s'adapter aux cas linéaires comme aux cas non linéaires tandis que les méthodes de régression requièrent d'identifier au préalable la structure de modèle adaptée à partir des données observées. Ce qui, excepté dans le cas de fonctions relativement simple (affine, polynôme, *etc.*) peut s'avérer très complexe. Le krigeage et la méthode des *splines* ont généralement des performances très proches. **Toutefois l'approche stochastique du krigeage est un plus car elle permet d'associer à chaque valeur prédite une incertitude de prédiction locale qui dépend de la topologie des données d'apprentissage.** Par la suite, nous avons choisi d'utiliser le krigeage comme méthode d'interpolation pour la modélisation du PT de la coupe gazole et du VI de la coupe huile.

Chapitre 5. Bases de données

L'objet de ce chapitre est de présenter les bases de données qui ont été utilisées pour l'étude du PT de la coupe gazole et du VI de la coupe huile. Le processus d'archivage des données issues d'expérimentations pilotes et de collectes des échantillons est préalablement décrit.

5.1 Processus d'expérimentation et archivage des données

Dans le but d'étudier les performances des catalyseurs mis en jeu lors du procédé d'HCK (paragraphe 1.2.2) et d'optimiser les conditions opératoires, des expérimentations sont effectuées au sein d'IFP Energies Nouvelles. Plusieurs types d'expérimentation sont réalisées :

1. Bilan HDT : le catalyseur utilisé ne fait que de l'HDT
2. Bilan HCK : le catalyseur utilisé ne fait que du cracking. La charge utilisée a donc été hydrotraitée au préalable.
3. Bilan HDT+HCK : les 2 catalyseurs d'HDT et d'HCK en série sont utilisés.

Au cours de ces expérimentations (volume de catalyseur entre 20 et 100cc), l'effluent est prélevé puis distillé pour obtenir les coupes pétrolières d'intérêt (essence, kérosène, gazole, UCO, huile, … (paragraphe 1.2.2). Les propriétés d'usage (paragraphe 3.1) des coupes (effluent compris) sont ainsi mesurées. La structure d'archivage des données issues des expérimentations pour le procédé d'HCK est récapitulée sur la Figure 33. L'ensemble des mesures provenant d'une même condition opératoire (charge, température, pression, temps de contact) est répertorié sous forme de bilan. Une base de données constituée de plusieurs bilans a ainsi été créée. Par la suite on désignera par bilan HDT, bilan HCK et bilan HDT+HCK tout ensemble de mesures issues respectivement d'un test HDT, d'un test HCK et d'un test HDT+HCK.

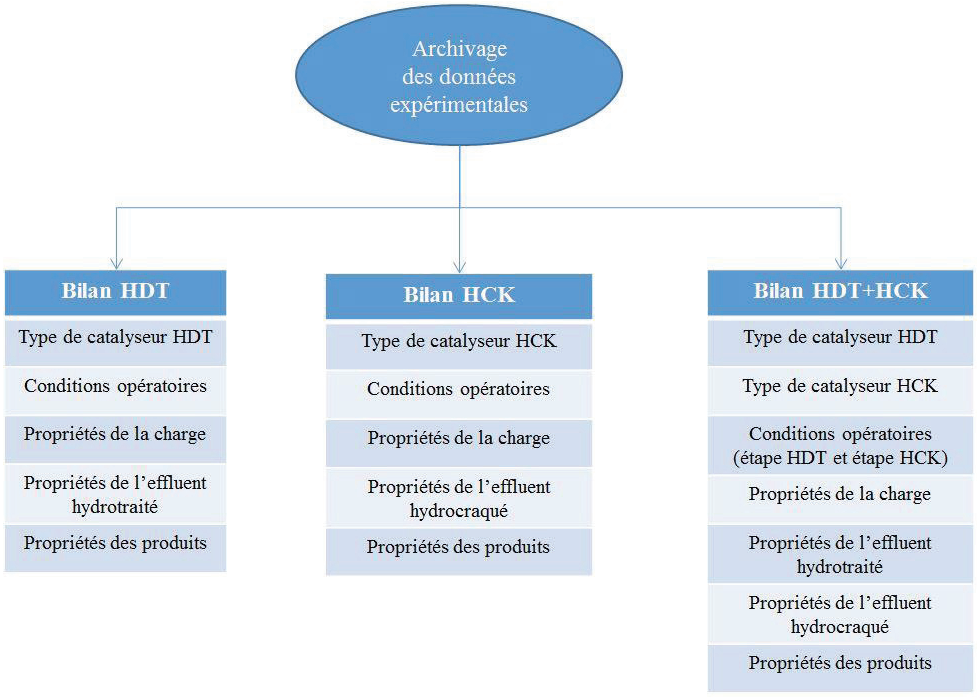


Figure 33 : Structure de l'archivage des données expérimentales pour le procédé d'hydrocraquage

5.1.1 Conditions expérimentales

5.1.1.1 Catalyseurs d'hydrotraitement

Les bilans exploités proviennent d'expérimentations qui ont été réalisées en faisant varier les conditions opératoires (température, pression, *etc.*) et en utilisant des catalyseurs différents. Pour les tests d'HDT (test HDT ou HDT+HCK), deux catalyseurs notés respectivement HDT_A et HDT_B, ainsi qu'un troisième catalyseur résultant de l'empilement (*stacking*) de HDT_A et HDT_B que nous noterons HDT_{A+B} ont été utilisés. Les catalyseurs d'HDT sont généralement constitués de métaux tels que le Molybdène (Mo) ou le Tungstène (W), associées au Cobalt (Co) ou au Nickel (Ni) et supporté par de l'alumine ou de la silice alumine [10,131].

5.1.1.2 Catalyseurs d'hydrocraquage

Pour les étapes de craquage (test HCK ou HDT+HCK), 4 catalyseurs différents ont été utilisés : les catalyseurs notés respectivement HCK_A, HCK_B qui présentent des caractéristiques distinctes (le second est plus actif et plus sélectif que le premier) ; les 2 catalyseurs résultant de leur empilement que nous noterons HCK_{A+B} et HCK_{B+A} (en fonction de l'ordre d'empilement).

Il est important de noter que les catalyseurs HCK_A et HCK_B n'utilisent pas la même zéolithe. Le catalyseur HCK_A est plus craquant mais moins isomérisant que le catalyseur HCK_B. Les propriétés des coupes sont également très différentes (meilleur VI et moins bon PT pour HCK_A).

5.1.1.3 Conditions opératoires

Les plages de variations des différentes conditions opératoires fixées lors des expérimentations sont listées dans le Tableau 15.

Le taux de conversion (X_{370}) a été défini au paragraphe 2.3.5.1. La Vitesse Volumétrique Horaire (VVH) est le terme utilisé pour désigner le débit de charge par volume de catalyseur. Elle s'exprime en h^{-1}. Enfin, PP_{H2} désigne la pression partielle en hydrogène.

Tableau 15 : Plage de valeurs pour les conditions opératoires utilisées dans lors des expérimentations

	Température (°C)	Pression totale (bar)	PP_{H2} (bar)	VVH (h^{-1})	X_{370} (%)
Réacteur 1 (HDT)	350 – 420	90 – 160	85 – 155	0,5 – 4,0	5 – 40
Réacteur 2 (HCK)	350 – 420	90 – 160	85 – 155	1,0 – 4,0	20 – 95

5.1.2 Charges

5.1.2.1 Distillats sous vide

Les bilans expérimentaux qui ont été exploités pour cette étude sont issus de 15 DSV différents. Les origines géographiques et les caractéristiques de ces charges sont précisées dans le Tableau 16.

5.1.2.2 Distillats sous vide prétraités

Lors d'expérimentation sur un catalyseur de cracking (test HCK), des charges dites prétraitées sont utilisées. Ces charges sont issues de l'HDT de DSV réalisé soit sur le catalyseur HDT_A, soit sur le catalyseur HDT_B, soit sur le catalyseur résultant de leur empilement (HDT_{A+B}) et dans des conditions opératoires très diverses (paragraphe 5.1.1.3). Les DSV d'origine des charges prétraitées ainsi que certaines de leurs caractéristiques sont précisés dans le Tableau 17.

Tableau 16 : Origine géographique et propriétés d'usage des distillats sous vide utilisées pour la production des échantillons analysés

DSV	Origine géographique et/ou type	Masse volumique à 15°C (g/cm^3)	% m/m Azote (ppm)	%m/m Soufre (ppm)	Intervalle de distillation (°C)	Viscosité à 70°C (cSt)	Viscosité à 100°C (cSt)
VGO A	Chine SR	0,8830	1015	1255	321 – 623	15,97	7,51
VGO B	Chine	0,8974	1190	10409	348 – 584	82,25	12,55
VGO C	Forcados	0,9429	1655	4170	308 – 601	41,00	13,60
VGO D	Iranian	0,9375	1300	28743	321 – 615	124,74	11,11
VGO E	Husky	0,9580	1080	30300	215 – 565	126,70	25,20
VGO F	Arabian Light	0,9237	835	24100	337 – 568	16,72	7,32
VGO G	Iranian Light	0,9563	1818	30232	317 – 729	110,50	31,94
VGO H	Oural SR	0,9234	1745	17711	285 – 605	72,64	7,65
VGO I	Arabian Heavy	0,9439	1255	29750	346 – 629	41,37	14,48
VGO J	HVO+HCGO	0,9849	4240	33200	286 – 643	339,20	14,00
VGO K	SR	0,9346	1755	22375	337 – 632	25,24	9,73
VGO L	Mélange DSV/HCGO 75/25	0,9393	2315	23300	246 – 610	20,17	8,13
VGO M	Hoil+Oural	0,9314	2160,00	10777	292 – 606	81,87	8,06
VGO N	US	0,9208	1160,00	2974	158 – 581	32,07	4,69
VGO P	Mélange Arabian Light/Basrah (Irak) 85/15	0,9284	1395,00	18921	316 – 624	133,24	11,93

Tableau 17 : Propriétés d'usage et origine des charges prétraitées

DSV d'origine	Label	Masse volumique à 15°C (g/cm^3)	% m/m Azote (ppm)	%m/m Soufre (ppm)	Intervalle d'ébullition (°C)	Viscosité à 70°C (cSt)	Viscosité à 100°C (cSt)
VGO B	VGO_HDT B	0,8741	2	0	160 – 562	14,61	6,84
VGO C	VGO_HDT C	0,9009	27	25	164 – 568	11,76	5,46
VGO E	VGO_HDT E	0,8948	8	86	169 – 513	76,89	26,24
VGO F	VGO_HDT F	0,8635	1	5	101 – 561	6,21	3,78
VGO G	VGO_HDT G	0,8878	15	72	178 – 631	17,61	7,88
VGO I	VGO_HDT I	0,8787	5	54	110 – 606	10,89	5,44
VGO K	VGO_HDT K0	0,8851	8	14	130 – 600	11,05	5,36
	VGO_HDT K1	0,8880	26	40	137 – 604	12,22	5,78
	VGO_HDT K2	0,8955	95	223	154 – 607	13,18	6,09
	VGO_HDT K3	0,9005	288	776	219 – 558		
	VGO_HDT K4	0,9011	302	875	166 – 612	14,80	6,64

5.2 Echantillons collectés

Dans ce paragraphe, nous proposons une présentation globale des échantillons de gazole et d'huile qui ont été collectés puis analysés par GC×GC, puis par RMN du ^{13}C.

5.2.1 Echantillons de gazole

Pour effectuer les travaux de compréhension moléculaire du PT, 40 échantillons de gazole ont été collectés. Ils sont issus de tests HCK réalisés sur 6 charges prétraitées et 4 catalyseurs différents (HCK_A, HCK_B, HCK_{B+A} et HCK_{A+B}). La répartition de ces échantillons par charge d'origine ainsi que des plages de valeurs de données expérimentales relatives à la production des échantillons sont précisées dans le Tableau 18. Pour chacun de ces échantillons de gazole, le point de trouble a été mesuré selon la norme NF EN ISO 23015 [168] avec un domaine compris entre -7 et -34°C. Ils ont ensuite été analysés par GC×GC (paragraphe 3.2.1) puis par spectroscopie RMN du ^{13}C (paragraphe 3.2.2).

Tableau 18 : Répartition des échantillons de gazole collectés par charge prétraitée d'origine

Charge	Nombre d'échantillons	$X_{370,min} - X_{370,max}$ (%)	$VVH_{min} - VVH_{max}$ (h^{-1})	$PT_{min} - PT_{max}$ (°C)
VGO_HDT I	20	70 – 95	1,5 – 3,0	-18,0 – -34,0
VGO_HDT K0	1	40,5 – 40,5	1,0 – 1,0	-18,0 – -18,0
VGO_HDT K1	4	43,0 – 85,5	1,5 – 1,5	-14,0 – -16,0
VGO_HDT K2	7	33,4 – 98,0	1,0 – 3,5	-18,0 – -27,0
VGO_HDT K4	6	31,0 – 91,2	1,0 – 3,5	-12,0 – -22,0
VGO_HDT F	2	26,0 – 85,5	1,0 – 1,0	-7,0 – -10,0
Total	**40**	**26,0 – 98,0**	**1,0 – 3,5**	**-7,0 – -34,0**

En dehors des échantillons présentés ci-dessus, des données issues d'analyse GC×GC de 6 autres échantillons de gazole ont été recueillies. Ces données proviennent d'un projet mené en interne à IFPEN. **Les 6 échantillons dont sont issues ces analyses ont été produits sur des catalyseurs différents de ceux qui ont été présentés dans ce chapitre**. Il nous a semblé judicieux de les utiliser pour évaluer la consistance des modèles développés à partir des données GC×GC. Les caractéristiques de ces échantillons sont données dans le Tableau 19.

Tableau 19 : Caractéristiques des échantillons analysés pour l'évaluation des modèles de prédiction du PT à partir de données GC×GC

Echantillon	Charge	PT(°C)
a	VGO_HDT X	-17
b	VGO_HDT X	-24
c	VGO_HDT K0	-21
d	VGO_HDT X	-16
e	VGO_HDT X	-29
f	VGO_HDT K0	-19

5.2.2 Echantillon d'huile

Pour les travaux de compréhension moléculaire du VI de la coupe huile, 31 échantillons issus exclusivement de tests HCK ont été collectés. Ces tests ont été effectués sur 5 charges prétraitées et 4 catalyseurs différents (HCK_A, HCK_B, HCK_{B+A} et HCK_{A+B}). La répartition de ces échantillons par charge (prétraitée) est donnée dans le

Tableau 20. Pour chacun de ces échantillons d'huile, le VI a été mesuré selon la norme NF EN ISO 2909 avec un domaine compris entre 92 et 125. Ces échantillons ont été analysés par GC×GC, puis par RMN du ^{13}C

Tableau 20 : Répartition des échantillons d'huile collectés par charge de provenance

Charge	Nombre d'échantillons	$X_{370,min} - X_{370,max}$ (%)	$VVH_{min} - VVH_{max}$ (h^{-1})	$VI_{min} - VI_{max}$ (°C)
VGO_HDT F	1	26,0 – 26,0	1,0 – 1,0	119 – 119
VGO_HDT I	17	70,0 – 95,0	1,5 – 3,0	102 – 124
VGO_HDT K1	3	43,0 – 85,5	1,5 – 1,5	102 – 125
VGO_HDT K2	5	31,0 – 91,2	1,0 – 3,5	92 – 109
VGO_HDT K4	5	39,0 – 98,0	1,0 – 2,0	92 – 109
Total	**31**	**26,0 – 98,0**	**1,0 – 3,5**	**92 – 125**

5.2.3 Echantillons produits à iso conditions opératoires

Pour comparer les impacts des différents catalyseurs utilisés pour la production des échantillons analysés au cours de cette étude, 4 échantillons d'huile et 4 échantillons de gazole ont été sélectionnés parmi les échantillons présentés au paragraphe précédent. Les échantillons d'huile, comme les échantillons de gazole sont issus de la même charge DSV (VGO I) et ont été produits sous des conditions opératoires (température, pression, *etc.*) très voisines mais en utilisant des catalyseurs différents. Ces échantillons sont représentés sur la Figure 34 d'une part dans le plan (X_{370},PT) pour les gazoles (Figure 34a) et d'autre part dans le plan (X_{370},VI) pour les huiles (Figure 34b). Chaque symbole renvoie au

catalyseur utilisé pour la production de l'échantillon : carré pour HCK_A, rond pour HCK_B, triangle pour HCK_{B+A} et losange pour HCK_{A+B}. Les barres d'erreur horizontales représentent les incertitudes liées au calcul de la conversion tandis que les barres verticales font référence à l'IC de la méthode standard de mesure de la propriété VI et PT [169].

La Figure 34a illustre une variation significative du PT suivant le catalyseur. En effet, le catalyseur HCK_A (ayant un pouvoir craquant plus fort) fournit des échantillons de gazole de plus haut PT que le ceux produit avec le catalyseur HCK_B malgré des conditions opératoires proches. *A contrario* on peut noter sur la Figure 34b que les échantillons d'huile produits avec le catalyseur HCK_B ont un VI plus bas que ceux qui ont été obtenus avec le catalyseur HCK_A. Les échantillons d'huile produits sur un enchainement de ces deux catalyseurs (HCK_{A+B} et HCK_{B+A}) ont quant à eux des valeurs PT et de VI intermédiaires relativement proches.

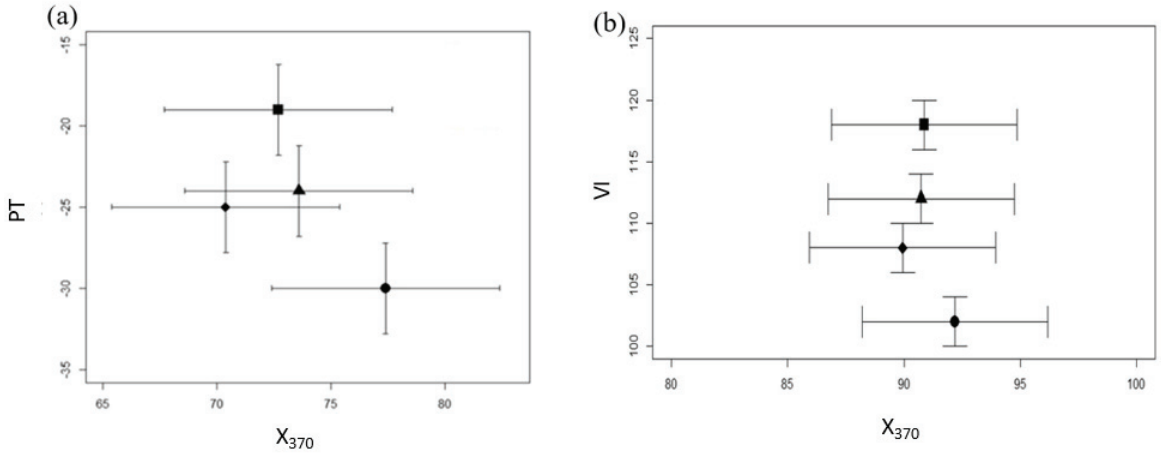


Figure 34 : Echantillons produits à iso conditions opératoires a) Evolution du PT en fonction du taux de conversion ; b) Evolution du VI en fonction du taux de conversion pour les échantillons d'huile produits à iso conditions opératoires ; catalyseurs : ■ → HCK_A ; ● → HCK_B ; ▲ → HCK_{B+A} ; ◆ → HCK_{A+B} ;

5.3 Bases de données – bilans archivés

Des données issues de bilans archivés ont été recueillies pour le développement de modèle de prédiction du PT de la coupe gazole et du VI de la coupe huile à partir de propriétés de base des coupes pétrolières (voir Chapitre 8). Les échantillons de coupes qui ont été analysés ne sont pas toujours conservés physiquement, ce qui explique les différences entre les bases de données utilisées pour ces travaux.

5.3.1 Base de données pour la prédiction du PT de la coupe gazole

Une base de données de 57 bilans archivés contenant les mesures de PT des coupes gazoles correspondantes a été créée. Ces données sont issues soit de bilans HCK, soit de bilans HDT+HCK. La répartition des gazoles par charge d'origine est donnée dans le Tableau 21.

Tableau 21 : Répartition des bilans de gazole par charge de provenance

Type de bilan	Charge	Nombre de bilans	$PT_{min} - PT_{max}$ (°C)
HDT+HCK	VGO K	11	-16 – -27
HDT+HCK	VGO M	6	-17 – -27
HDT+HCK	VGO N	6	-25 – -39
HDT+HCK	VGO P	1	-21 – -21
HCK	VGO_HDT B	3	-25 – -31
HCK	VGO_HDT C	4	-19 – -25
HCK	VGO_HDT G	3	-24 – -30
HCK	VGO_HDT K0	9	-17 – -34
HCK	VGO_HDT K2	4	-20 – -32
HCK	VGO_HDT K3	2	-12 – -17
HCK	VGO_HDT K4	6	-18 – -28
HCK	VGO_HDT I	1	-34 – -34
Total	**12**	**56**	**-16 – -39**

5.3.2 Bases de données pour la prédiction du VI des huiles

Pour le développement de modèles de prédiction du VI de la coupe huile, deux bases de données ont été créées. La première concerne des huiles produites par HDT des DSV. Elle contient des propriétés d'usage d'effluents hydrotraités et les VI des coupes huiles correspondantes. Ces données sont issues de 135 bilans HDT provenant de tests réalisés sur 8 DSV différents et dans des conditions opératoires diverses (Tableau 16).

La seconde base de données contient les propriétés d'usage d'effluents hydrocraqués et les VI des coupes huiles correspondantes. Ces données sont issues de 82 bilans HCK provenant de tests réalisés sur 7 charges DSV prétraités, dans des conditions opératoires diverses (Tableau 17).

La répartition des bilans par type et par charge d'origine est donnée dans le Tableau 22. Les intervalles de valeurs de VI des coupes huiles correspondantes sont également précisés.

Tableau 22 : Répartition des bilans HDT/HCK utilisés pour la prédiction du VI par charge de provenance

Base de données	Charge	Nombre de bilans	$VI_{min} - VI_{max}$ (°C)
Bilans HDT	VGO A	6	80 – 97
	VGO D	7	70 – 96
	VGO H	10	70 – 113
	VGO J	11	9 – 91
	VGO K	56	53 – 106,8
	VGO L	10	66 – 85
	VGO M	22	63 – 79
	VGO N	14	25 – 58
Total HDT	**8**	**135**	**9 – 113**
Bilans HCK	VGO_HDT F	1	103 – 103
	VGO_HDT I	20	102 – 126
	VGO_HDT K0	20	93 – 114
	VGO_HDT K1	17	96 – 131
	VGO_HDT K2	12	92 – 109
	VGO_HDT K3	3	87 – 105
	VGO_HDT K4	11	85 – 109
Total HCK	**7**	**82**	**85 – 131**

Partie 3 : Résultats et discussions

Nous avons présenté les bases de données qui vont être utilisées, les techniques d'analyse dont elles sont issues et leur mode d'acquisition, ainsi que les outils statistiques sur lesquels nous allons nous appuyer. Cette troisième partie a pour objectif de présenter et discuter les différents résultats obtenus. Ces derniers sont commentés à la lumière des précédentes études qui ont été effectuées sur des problématiques voisines. Cette partie est structurée comme suit : le Chapitre 6 présente les résultats de l'exploitation des données issues de la caractérisation des coupes pétrolières par GC×GC aussi bien dans le cas du PT de la coupe gazole que dans celui du VI de la coupe huile ; le Chapitre 7 revient sur les études analogues qui ont été menées à partir des données issues de la RMN du ^{13}C pour ces mêmes propriétés; enfin, dans le Chapitre 8 nous proposerons une comparaison des différents modèles prédictifs qui ont été développés à partir de propriétés de base de coupes pétrolières.

Chapitre 6. Compréhension moléculaire des propriétés produits par GC×GC

L'objet de ce chapitre est de présenter les études de compréhension moléculaire du PT de la coupe gazole et du VI de la coupe huile à partir des données issues de la GC×GC.

6.1 Etude du point de trouble de la coupe gazole

Dans ce paragraphe, nous présentons les résultats de l'exploitation des données issues des analyses GC×GC des échantillons de gazole pour la compréhension moléculaire du PT. Une analyse exploratoire des données est proposée en amont. Ensuite l'étude qui a été menée est présentée suivant deux approches : une première approche dite empirique qui consiste à observer des tendances *via* des représentations graphiques des échantillons caractérisés et qui est souvent utilisée au sein d'IFPEN ; une seconde approche par analyse multivariée appliquée aux données chromatographiques (paragraphe 3.2.1.2).

6.1.1 Analyse exploratoire des échantillons

Pour visualiser globalement l'ensemble des échantillons analysés, une ACP a été réalisée à partir des données chromatographiques.

La Figure 35 représente les scores des échantillons de gazole caractérisés par GC×GC sur les composantes (PC1, PC2) qui expriment 69% de la variance expliquée. Les valeurs de PT de référence de ces échantillons sont répertoriées par une échelle de couleur (bleu pour les PT bas et rouge pour les PT hauts) et les différents symboles renvoient aux charges DSV à partir desquelles ces échantillons ont été produits (Tableau 16). Les vecteurs de *loadings* (ou poids) des variables chromatographiques sur les axes PC1 et PC2 sont représentés sur la Figure 36 et sur la Figure 37 respectivement.

La répartition des échantillons sur (PC1, PC2) montre la présence d'un *cluster* (groupe de points) représentatif d'échantillons provenant essentiellement de la charge VGO_HDT I (Figure 35). Selon les vecteurs de *loading*s sur la composante PC1 (Figure 36), ces échantillons se démarquent des autres par une proportion relativement élevé de composés n-paraffiniques (C_{13} à C_{19}), isoparaffiniques (C_{13} à C_{19}) et mononaphténiques (C_{13} à C_{17}). A l'opposé, les échantillons issues de charges-filles VGO_HDT K qui présentent une forte variabilité suivant la composante PC1 (Figure 36) ont des proportions relativement importantes de composés aromatiques et dinaphténiques (C_{15} à C_{23}). Enfin, on remarque que les échantillons qui proviennent de VGO_HDT F possèdent un score particulièrement élevé sur la composante PC2 qui traduit à la fois une forte présence de composés naphténiques et paraffiniques (C_7 à C_{11} et C_{21} à C_{23}) et une faible présence de composés polyaromatiques comme illustré sur la Figure 37. L'ACP des données chromatographiques met en évidence une nette différence de composition entre les échantillons. Pour le groupe d'échantillons issus de la charge VGO_HDT I la

proportion relativement importante de composés paraffiniques et mononaphténiques peut être expliquée par le haut taux de conversion appliqué à cette charge (70 - 95% en 370^+, voir Tableau 18). Pour les autres échantillons (VGO_HDT K), les domaines de conversion sont plus faibles (40 – 90 % en 370^+, voir Tableau 18) impliquant une zone de PT plus étendue, augmentant de gauche à droite (suivant PC1) et de bas en haut (suivant PC2) (Figure 35). Rappelons qu'une conversion plus élevée va correspondre à une hydrogénation plus marquée des espèces aromatiques, un craquage des molécules et une isomérisation de ces dernières allant vers plus de n-paraffines et d'isoparaffines.

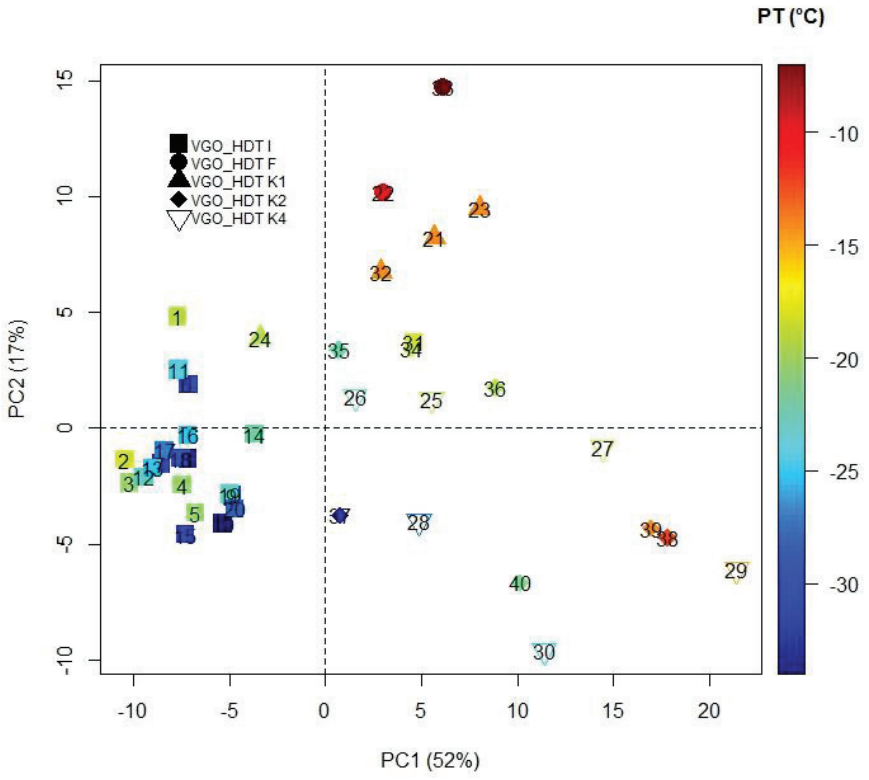


Figure 35 : Scores des échantillons de gazoles caractérisés par GC×GC dans le plan factoriel (PC1, PC2)

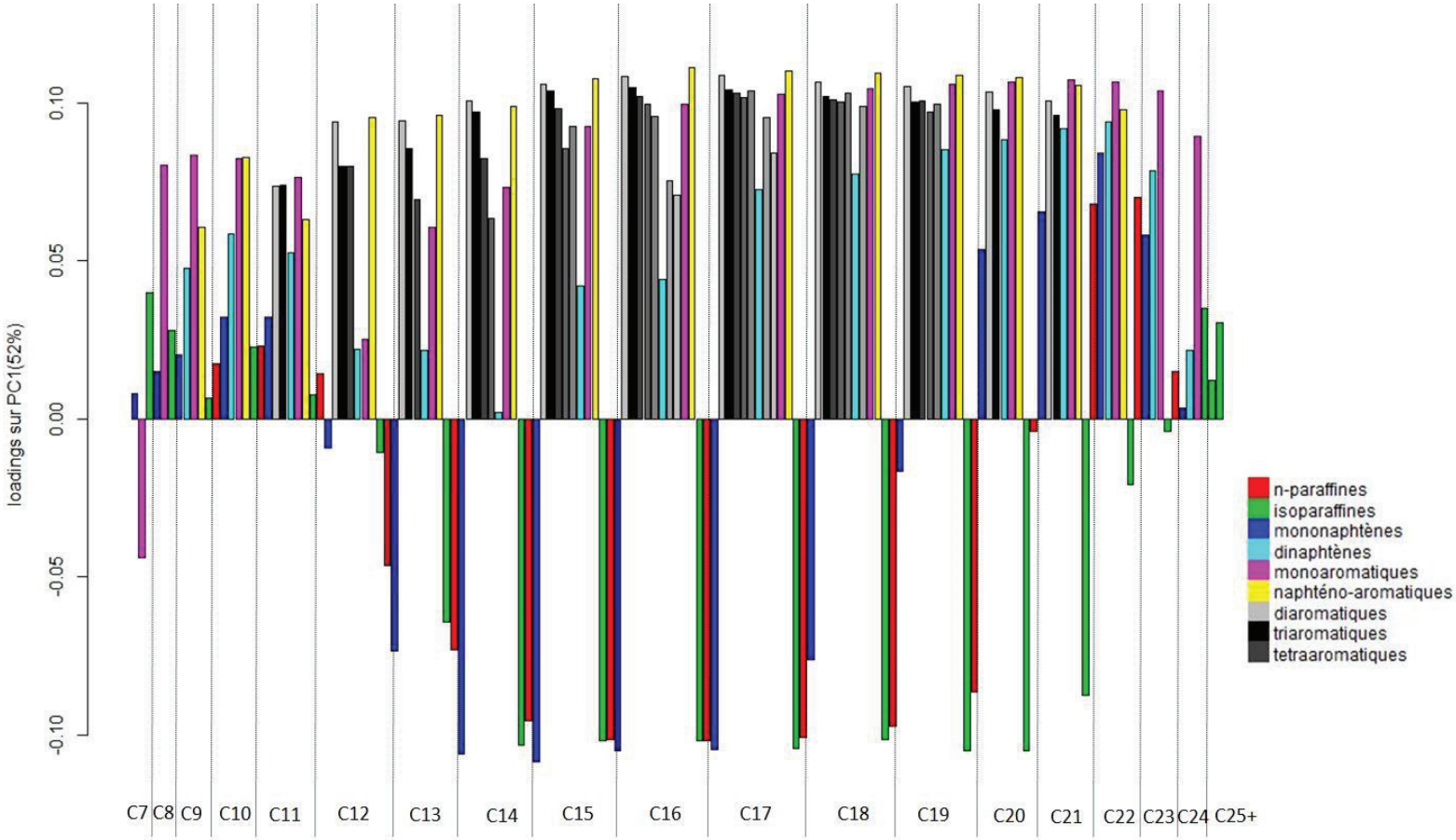


Figure 36 : *Loadings* des variables chromatographiques sur la composante PC1

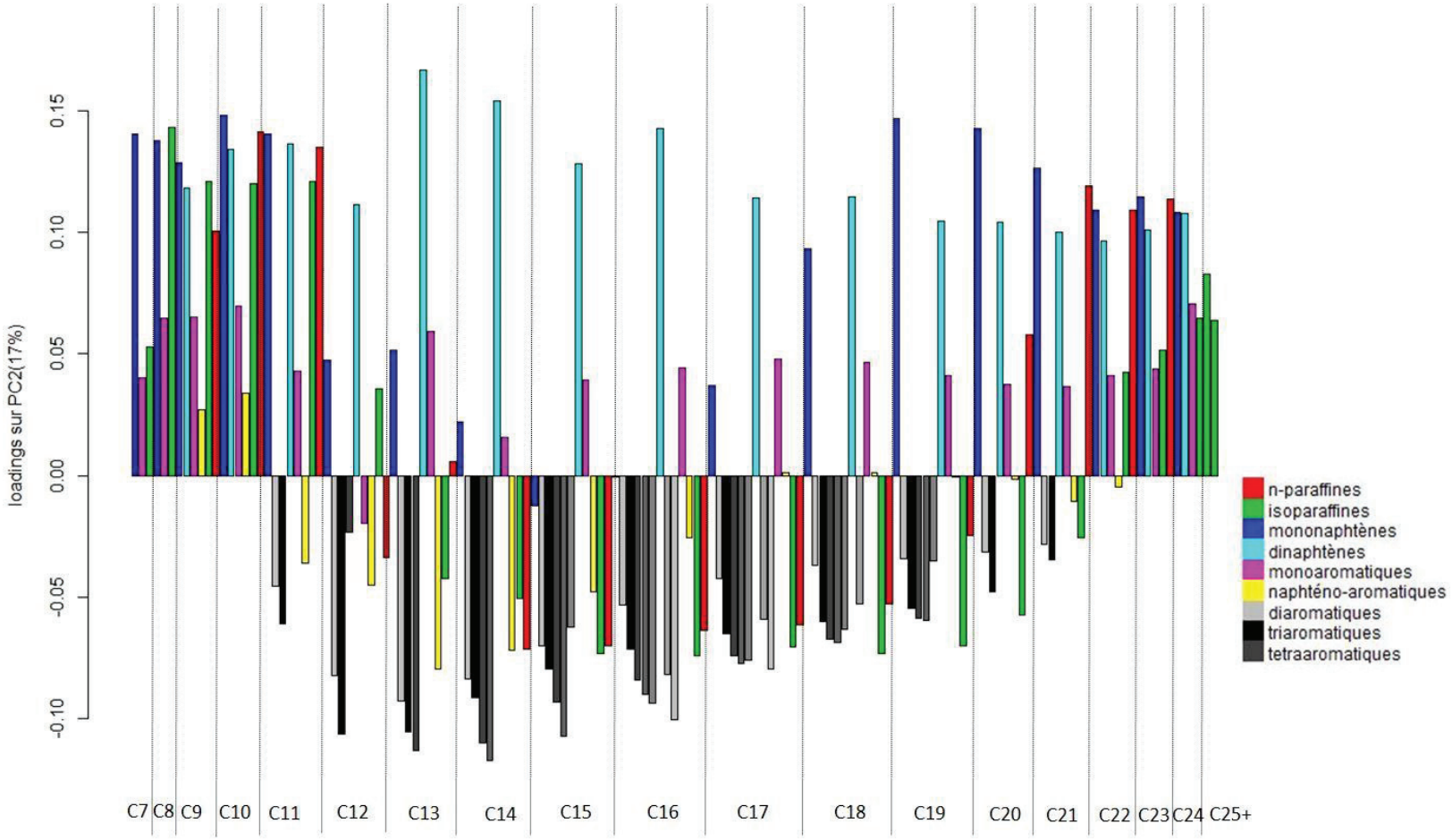


Figure 37 : *Loadings* des variables chromatographiques sur la composante PC2

6.1.2 Approche empirique

L'étude bibliographique concernant les propriétés à froid de la coupe gazole a mis en évidence un lien entre le PT et la proportion de n-paraffines contenue dans la coupe (paragraphe 2.4). Certaines de ces études ont également soulignées un rôle des isoparaffines dans le ralentissement du processus de cristallisation des n-paraffines (ce qui a pour conséquence de baisser le PT). L'approche proposée ici consiste en quatre points :

1. Observer l'évolution du PT en fonction de la teneur en divers composés, principalement ceux qui sont mentionnés dans l'étude bibliographique.
2. Rechercher des dépendances en identifiant si possible la forme des nuages de points observé,
3. Définir un modèle de prédiction du PT à partir de la concentration des composés qui montrent une influence nette sur le PT.
4. Evaluer les performances du modèle pour valider la pertinence des composés retenues.

La Figure 38 représente l'évolution du PT en fonction de la concentration en n-paraffines obtenue par GC×GC sur l'ensemble des échantillons de la base de données. On remarque des tendances différentes suivant la charge DSV d'origine : pour les échantillons issus de la charge VGO_HDT I (entourés en bleu sur la Figure 38), le PT tend à augmenter avec la concentration en n-paraffines ; Il en est de même pour les échantillons produits à partir de VGO_HDT F (entourés en vert sur la Figure 38). Par contre pour les échantillons issus de charges filles VGO_HDT K (entourés en rouge sur la Figure 38), le PT varie de façon importante quasi-indépendamment de la concentration en n-paraffines, au demeurant relativement faible (sensiblement 4% poids). Il apparait donc important d'observer en complément de la concentration en n-paraffines, les distributions en nombre de carbones de ces dernières qui peuvent influer sur la valeur du PT.

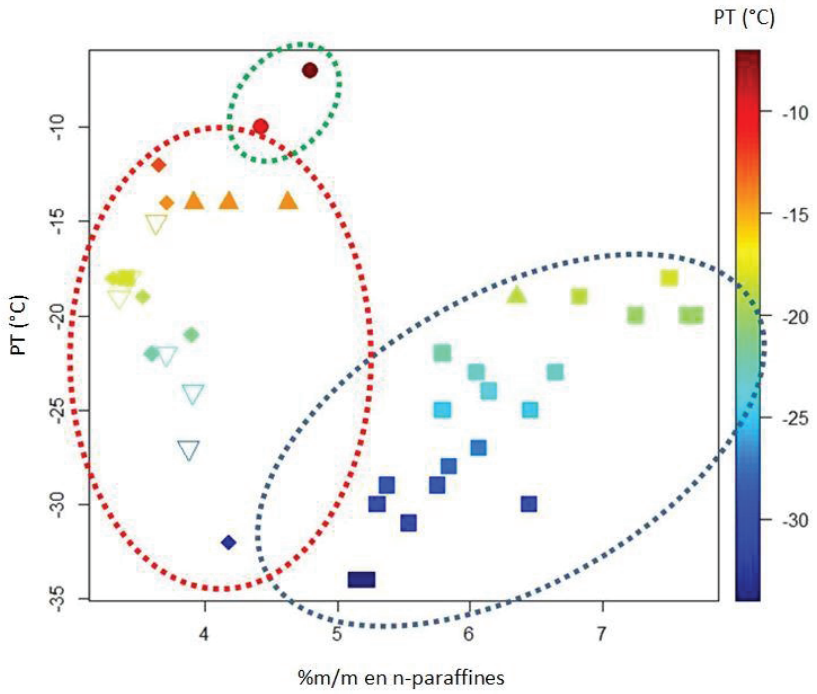


Figure 38 : Evolution du PT en fonction de la proportion en n-paraffines contenue dans les échantillons de gazole ; Les DSV d'origine sont représentés par des symboles : ■ → VGO_HDT I ; ● → VGO_HDT F ; ▲ → VGO_HDT K1 ; ◆ → VGO_HDT K2 ; ▼ → VGO_HDT K4

En conséquence, l'évolution du PT en fonction de la teneur des différentes espèces de n-paraffines a ainsi été étudiée, l'objectif étant d'obtenir une dépendance commune à l'ensemble des échantillons de la base de données. A noter que pour les n-paraffines possédant moins de 20 atomes de carbone, les graphes n'ont fourni aucun résultat concluant (voir Annexe I). Cette observation rejoint les conclusions faites par Krishna *et al.* [67] quant à l'indépendance du PT par rapport à la teneur en n-paraffines C_{15-}. Nous avons représenté sur la Figure 39 l'évolution du PT en fonction des n-paraffines C_{20} (Figure 39a), C_{21} (Figure 39b), C_{22} (Figure 39c) et C_{23} (Figure 39d). Ces dernières constituent les plus longues n-paraffines qui ont été quantifiées dans les 40 échantillons de gazole analysés. On note que plus le nombre de carbones augmente, plus des tendances quasi-linéaires peuvent être observées dans le cas des n-paraffines de longueur C_{20} à C_{22}. Pour la n-paraffine C_{23}, on observe une quasi-indépendance de l'évolution du PT (Figure 39d). A noter le comportement singulier du gazole de plus haut PT (-7°C) issu de la charge VGO_HDT F pour une concentration en n-Tricosane (C_{23}) la plus élevée (>0,25% poids).

Des tendances linéaires globales ne sont donc observées que dans le cas des n-paraffines de longue chaîne (C_{20} à C_{22}). Pour la n-paraffine C_{23}, il semble que la teneur soit en général trop faible pour influer sur le PT (excepté pour l'échantillon de plus haut PT). **Le couple concentration et distribution**

en taille des n-paraffines a une influence prépondérante sur le PT, en adéquation avec les conclusions de l'étude bibliographique.

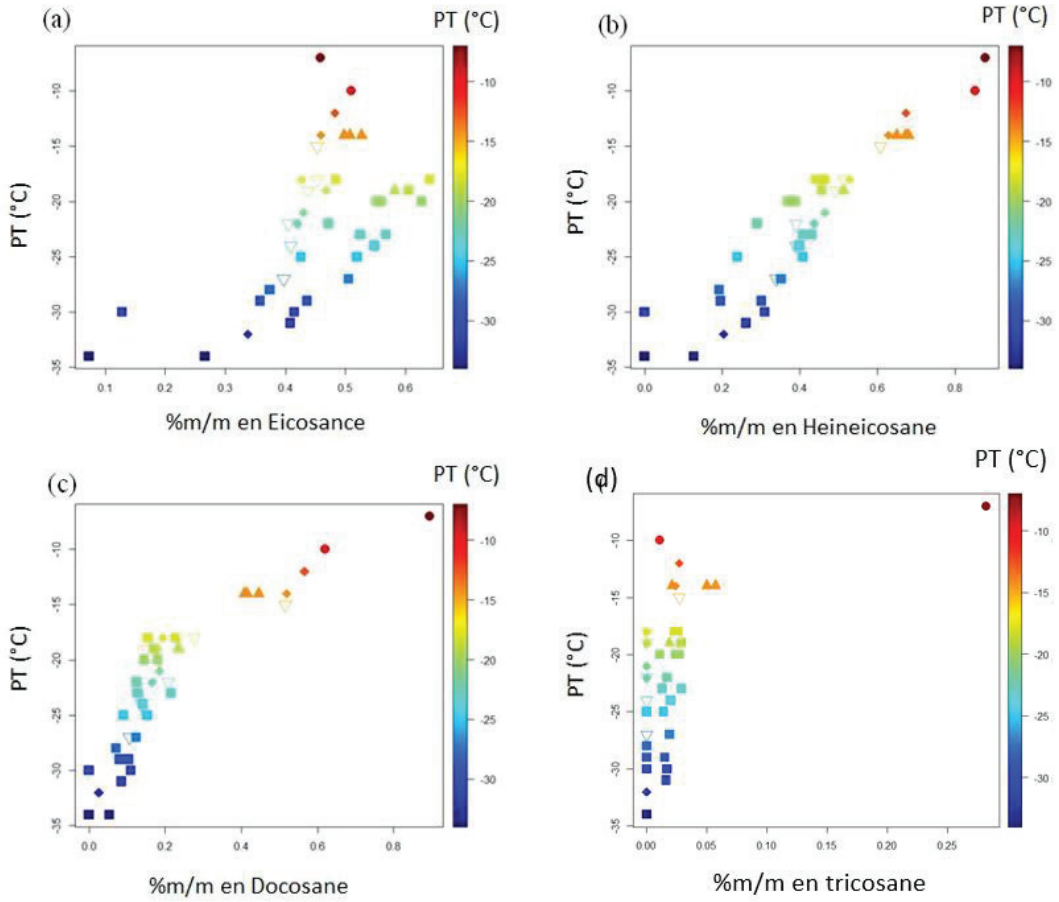


Figure 39 : Evolution du PT en fonction de la teneur en : (a) n-Eicosane (n-C_{20}) ; (b) n-Heneicosane (n-C_{21}) ; (c) n-Docosane (n-C_{22}) ; (d) n-Tricosane (n-C_{23}) ; Les DSV d'origine sont représentés par des symboles : ■ → VGO_HDT I ; ● → VGO_HDT F ; ▲ → VGO_HDT K1 ; ◆ → VGO_HDT K2 ; ▼ → VGO_HDT K4

L'influence des isoparaffines sur le PT a également été étudiée. Les évolutions du PT en fonction de la teneur en isoparaffines, puis en fonction du rapport teneur en n-paraffines sur teneur en iso paraffines sont illustrées sur la Figure 40. Dans le premier cas on observe une diminution du PT quand la teneur en isoparaffines augmente pour les échantillons issus de la charge VGO_HDT K (Figure 40a). Une dépendance plus nette est visible dans le second cas (Figure 40b), le PT augmentant globalement avec le rapport des teneurs en n-paraffines et isoparaffines.

Ces observations montrent que la teneur en isoparaffines est un facteur de diminution du PT. Elles mettent par ailleurs en évidence d'une part que l'effet de la teneur en isoparaffines sur le PT n'est

pas linéaire sur toute la gamme, et d'autre part que cet effet est indissociable de la présence des n-paraffines. **Ce constat est en adéquation avec les conclusions de la littérature concernant l'effet de ralentissement causé par les isoparaffines dans le processus de cristallisation des n-paraffines [84].**

Par la suite, nous avons effectué des études analogues pour les autres familles de composés en traçant les évolutions du PT en fonction des teneurs en monoaromatiques, en diaromatiques, en mononaphtènes ou encore en dinaphtènes (graphes disponibles en Annexe J), mais ces travaux n'ont pas abouti à des conclusions intéressantes.

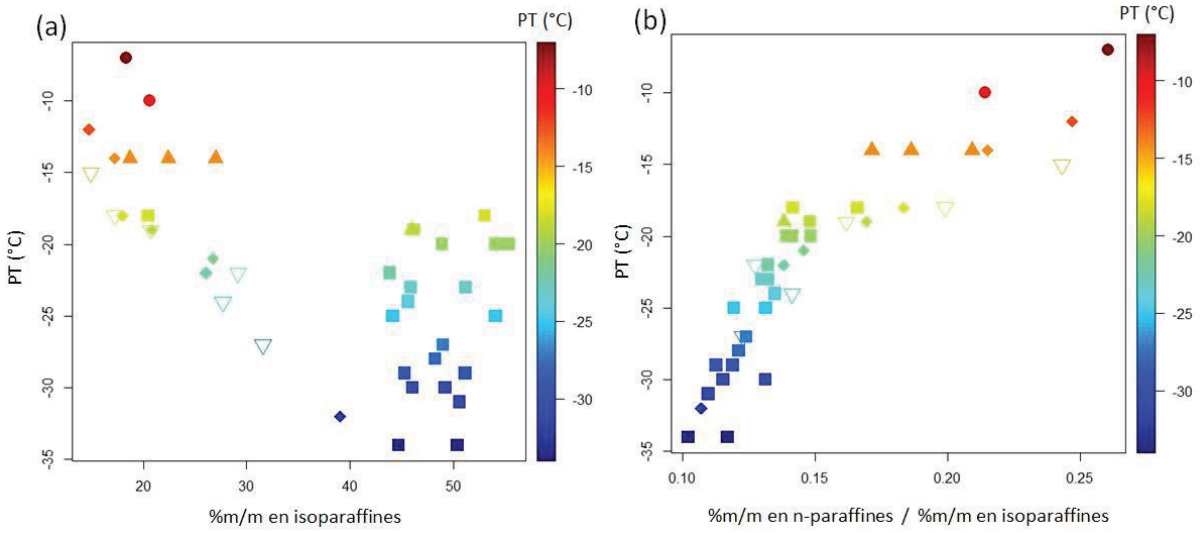


Figure 40 : Evolution du PT ; (a) en fonction de la teneur en isoparaffines ; (b) en fonction du rapport des teneurs en n-paraffines et isoparaffines ; Les DSV d'origine sont représentés par des symboles : ■ → VGO_HDT I ; ● → VGO_HDT F ; ▲ → VGO_HDT K1 ; ◆ → VGO_HDT K2 ; ▼ → VGO_HDT K4

A partir des observations ci-dessus concernant l'importance des n et isoparaffines, un modèle RLM pour la prédiction du PT a été développé. Les descripteurs sont :

- Les teneurs en n-paraffines C_{19}, C_{20}, C_{21}, C_{22} et C_{23}
- La teneur globale en isoparaffines

La base d'apprentissage est constituée des 40 échantillons de gazole que nous avons collectés (Tableau 18). La base de test est constituée des 6 échantillons indépendants présentés au paragraphe 5.2.1 (Tableau 19). Les graphes de parité du modèle obtenu sur la base d'apprentissage et sur la base de test sont illustrés sur la Figure 41. Les droites en pointillés verts et bleus délimitent les bornes de l'IC associé à la méthode de mesure du PT. Les statistiques du modèle sont données dans le Tableau 23. On observe une bonne qualité globale de la prédiction aussi bien sur les données d'apprentissage que sur les données

de test qui est confirmé par les statistiques. En effet, une erreur quadratique moyenne de prédiction (RMSEP) de 1,93 a été obtenue sur l'ensemble des données d'apprentissage et 77% de ces échantillons sont prédits avec une erreur inférieure à l'IC de la mesure. Sur la base de test, la valeur de la RMSEP est de 1,4°C et 100% des points sont prédits à l'intérieur de l'IC de référence. Bien que les données de test ne couvrent pas toute la gamme de PT, **ce résultat donne une indication quant à la bonne qualité du modèle et confirme donc l'influence des différents composés mis en jeu sur le PT de la coupe gazole.**

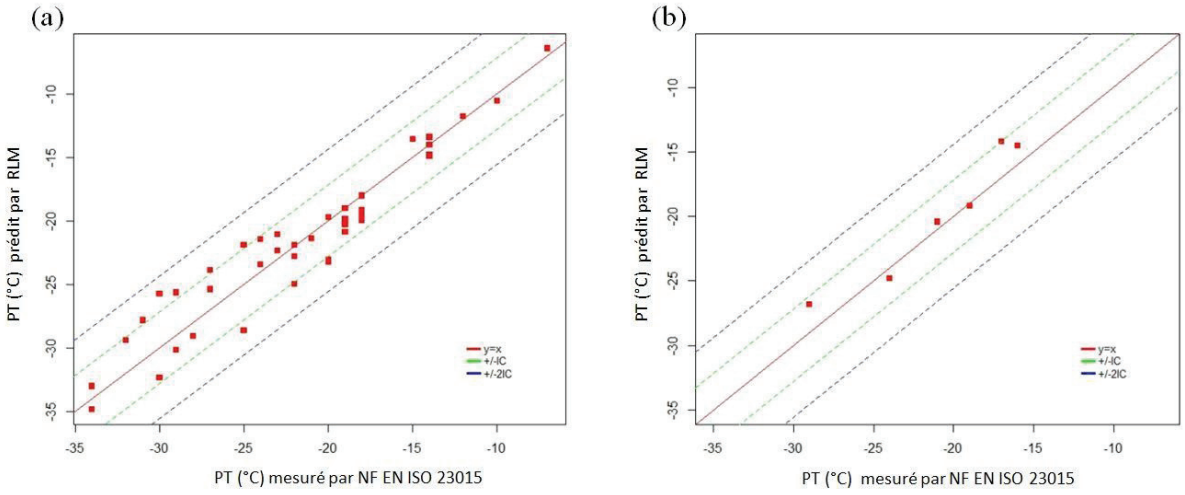


Figure 41 : Graphes de parités du modèle empirique ; (a) sur la base d'apprentissage ; (b) sur la base de test

Tableau 23 : Statistiques du modèle empirique de prédiction du PT de la coupe gazole à partir des données GC×GC

Base d'évaluation	RMSE* (°C)	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)	IC méthode NF ISO 23015 (°C)
Apprentissage	1,93	77,5	100	2,8
Test	1,4	100	100	

*→RMSEC pour la base d'apprentissage et RMSEP pour la base de test

L'approche empirique présentée au paragraphe précédent est intéressante. Elles requièrent cependant un grand nombre d'interprétations graphiques et ne tient pas compte des inter-corrélations entre les variables.

6.1.3 Régression multivariée pour la modélisation du PT de la coupe gazole

Pour s'affranchir de ces inconvénients, les données GC×GC ont été exploitées par analyse multivariée, conformément à la méthodologie définie (paragraphe 2.5.2).

6.1.3.1 Modèle de régression PLS

Un modèle de régression PLS a été développé pour vérifier la pertinence des informations contenues dans les données GC×GC (Figure 15) en vue d'expliquer le PT des échantillons de gazole. **Ces données ont été préalablement centrées et réduites.** La base d'apprentissage est toujours constituée des 40 échantillons de gazole que nous avons collectés (Tableau 18). La courbe d'évolution de la RMSECV est représentée sur la Figure 42. On observe une décroissance de la RMSECV en fonction du nombre de variables latentes jusqu'à un nombre de variables latentes égal à 15. Ce nombre de variables étant très élevé en comparaison aux données d'apprentissage (40), nous avons choisi de fixer le nombre de variables latentes à 10, nombre à partir duquel la RMSECV ne varie plus significativement.

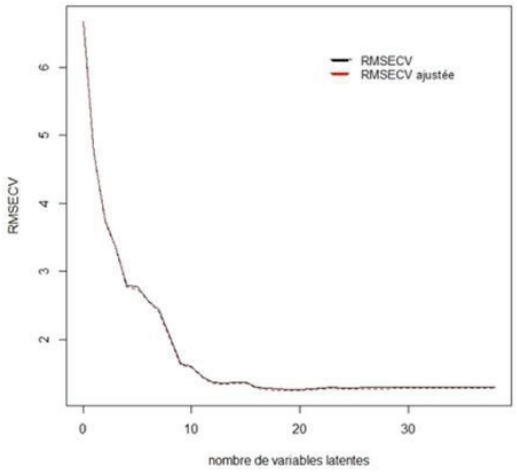


Figure 42 : Evolution de la RMSECV en fonction du nombre de variables latentes

Le modèle a été évalué sur la base d'apprentissage, puis sur des données de test identiques à celles qui ont été introduites pour le modèle RLM (paragraphe 6.1.2). Les graphes de parité obtenus sont représentés sur la Figure 43. Les statistiques du modèle sont précisées dans le Tableau 24. Une RMSEC de 0,56°C a été obtenue sur les données d'apprentissage et 100% des points sont prédits à l'intérieur de l'intervalle de confiance de la mesure. Les statistiques estimées sur la base de test sont moins bonnes (RMSEP de 2,2°C et 83% des points prédits à l'intérieur de l'IC de référence) mais tout à fait satisfaisante. Ces résultats confirment les résultats du modèle RLM. La différence RMSEC et RMSEP traduit un probable surparamétrage du modèle PLS. Une explication possible est que le nombre d'échantillons utilisés pour la phase d'apprentissage est insuffisant.

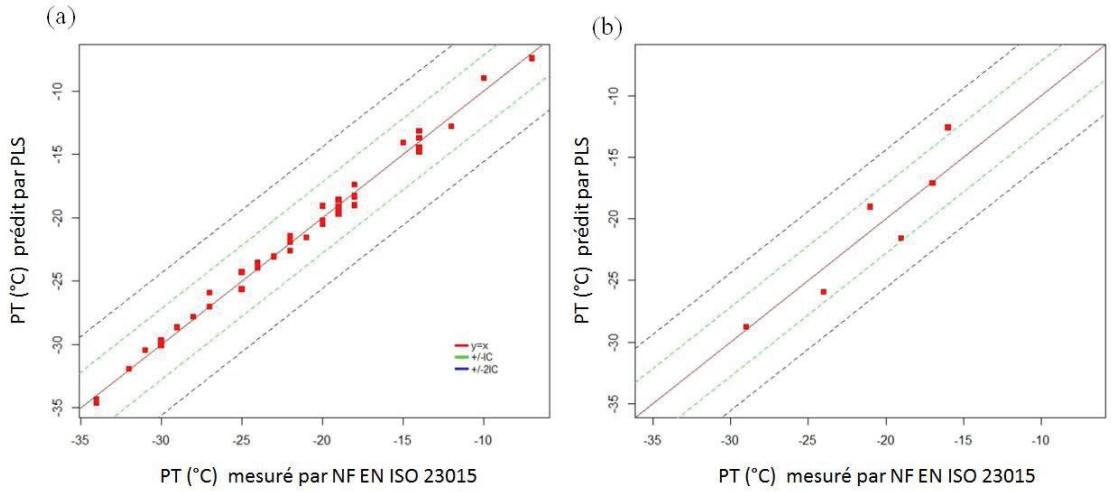


Figure 43 : Graphes de parités du modèle de prédiction du PT de la coupe gazole par régression PLS ; (a) sur la base d'apprentissage ; (b) sur la base de test

Tableau 24 : Statistiques du modèle de prédiction du PT de la coupe gazole par régression PLS appliquée aux données GC×GC

Base d'évaluation	RMSECV (°C)	RMSE* (°C)	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)	IC méthode NF ISO 23015 (°C)
Apprentissage	1,6	0,56	100	100	2,8
Test		2,3	83	100	

*→RMSEC pour la base d'apprentissage et RMSEP pour la base de test

6.1.3.2 Modélisation du PT par sparse PLS

Pour identifier ainsi les composés essentiels à la modélisation du PT, un modèle *sparse* PLS a été développé pour un nombre de variables latentes équivalent à celui du modèle PLS (les données d'apprentissage sont également identiques). Comme la *sparse* PLS est utilisée ici pour l'identification des facteurs prépondérants, nous avons choisi de ne pas l'évaluer sur la base de test. Ce choix se justifie par le fait qu'ici le nombre de variables latentes est imposé par le modèle PLS alors qu'il pourrait être fixé par validation croisée (en même temps que le coefficient de seuillage). L'évolution de la RMSECV en fonction de η (pour η compris entre 0 et 1) est représentée sur la Figure 44. On observe une valeur minimale de la RMSECV (1,20°C) pour $\eta = 0{,}86$.

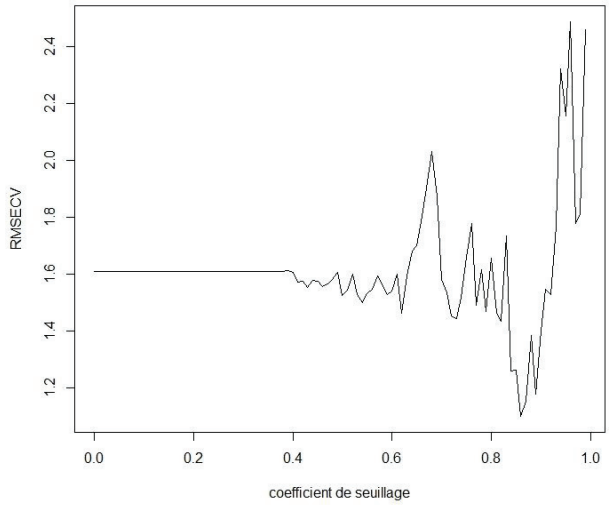


Figure 44 : Evolution de la RMSECV en fonction du coefficient de seuillage pour 10 variables latentes

Le graphe de parité du modèle *sparse* PLS correspondant sur la base d'apprentissage est représenté sur la Figure 45. On peut noter qu'il est très proche de celui du modèle PLS (Figure 43a). Ce constat est appuyé par les statistiques de performances précisées dans le Tableau 25. En effet, RMSEC de 0,48°C a été obtenue et de même que pour le modèle PLS 100% des échantillons d'apprentissage sont prédits avec une erreur inférieure à l'IC de la mesure de référence. Ces observations confirment l'intérêt de l'utilisation de la *sparse* PLS dans ce cas précis.

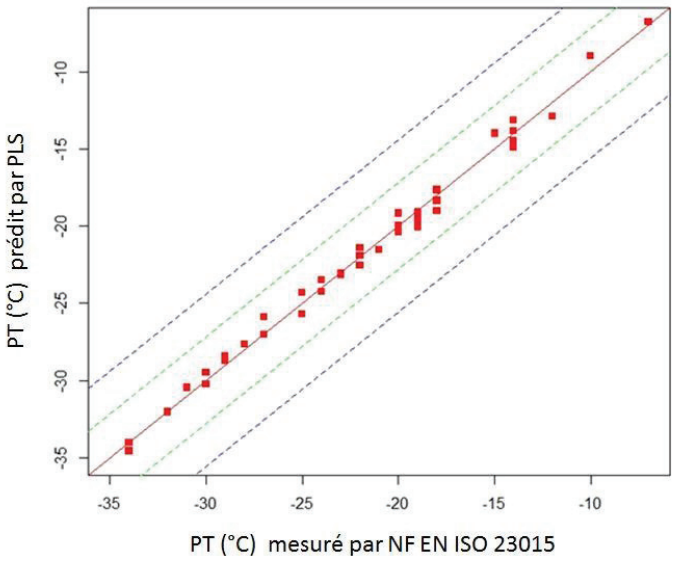


Figure 45 : Graphe de parité du modèle de prédiction du PT de la coupe gazole par *sparse* PLS appliquée aux données sur la base d'apprentissage

Tableau 25 : Statistiques du modèle de prédiction du PT de la coupe gazole par *sparse* PLS appliquée aux données GC×GC sur la base d'apprentissage

Modèle	Nbre variables latentes	Coefficients de seuillage	RMSECV (°C)	RMSEC (°C)	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
sparse PLS	10	0,86	1,20	0,48	100	100

6.1.3.3 Interprétation de la sparse PLS

La Figure 46 illustre les scores des échantillons de gazole caractérisés par GC×GC sur les deux premières variables latentes. Les vecteurs de *loadings* associés à ces composantes sont représentés respectivement sur la Figure 47 et la Figure 48. Deux tendances sont notables : pour les échantillons issus de VGO_HDT F et VGO_HDT K, le PT augmente suivant la composante *sparse* LV1 ; pour les échantillons produits à partir de VGO_HDT I, le PT augmente globalement suivant *sparse* LV2. Plus généralement, on remarque que les échantillons qui ont les PT les plus hauts sont ceux qui ont un score particulièrement élevé sur *sparse* LV1 et sur *sparse* LV2.

Les *loadings* associés aux points ayant un PT élevé montrent d'une part une forte teneur en n-paraffines C_{21} et C_{22}, en mononaphtènes et dinaphtènes C_{22} (Figure 47) et la présence moins significative des n-paraffines C_{18} et d'isoparaffines C_{21}. Les *loadings* associés aux échantillons ayant un PT bas ont majoritairement un score faible sur *sparse* LV1 et sur *sparse* LV2. Les *loadings* associés aux échantillons ayant un PT intermédiaire (autour de -20°C) ont soit un score moyen suivant les deux composantes, soit un score élevé suivant la composante *sparse* LV2. Cette dernière traduit d'une part une forte teneur en n-paraffines C_{18} à C_{22}, avec une prépondérance pour la n-paraffine C_{20} et d'autre part une faible teneur en aromatiques (Figure 48).

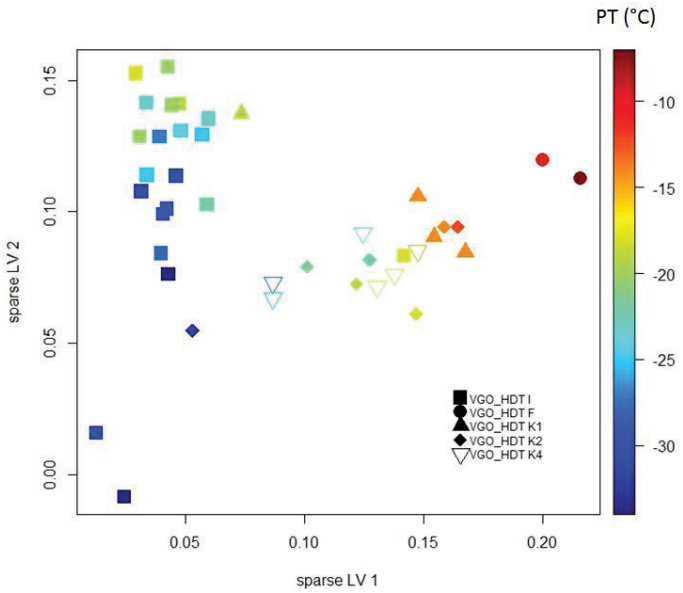


Figure 46 : Scores des échantillons de gazole caractérisés par GC×GC sur les deux premières composantes parcimonieuses

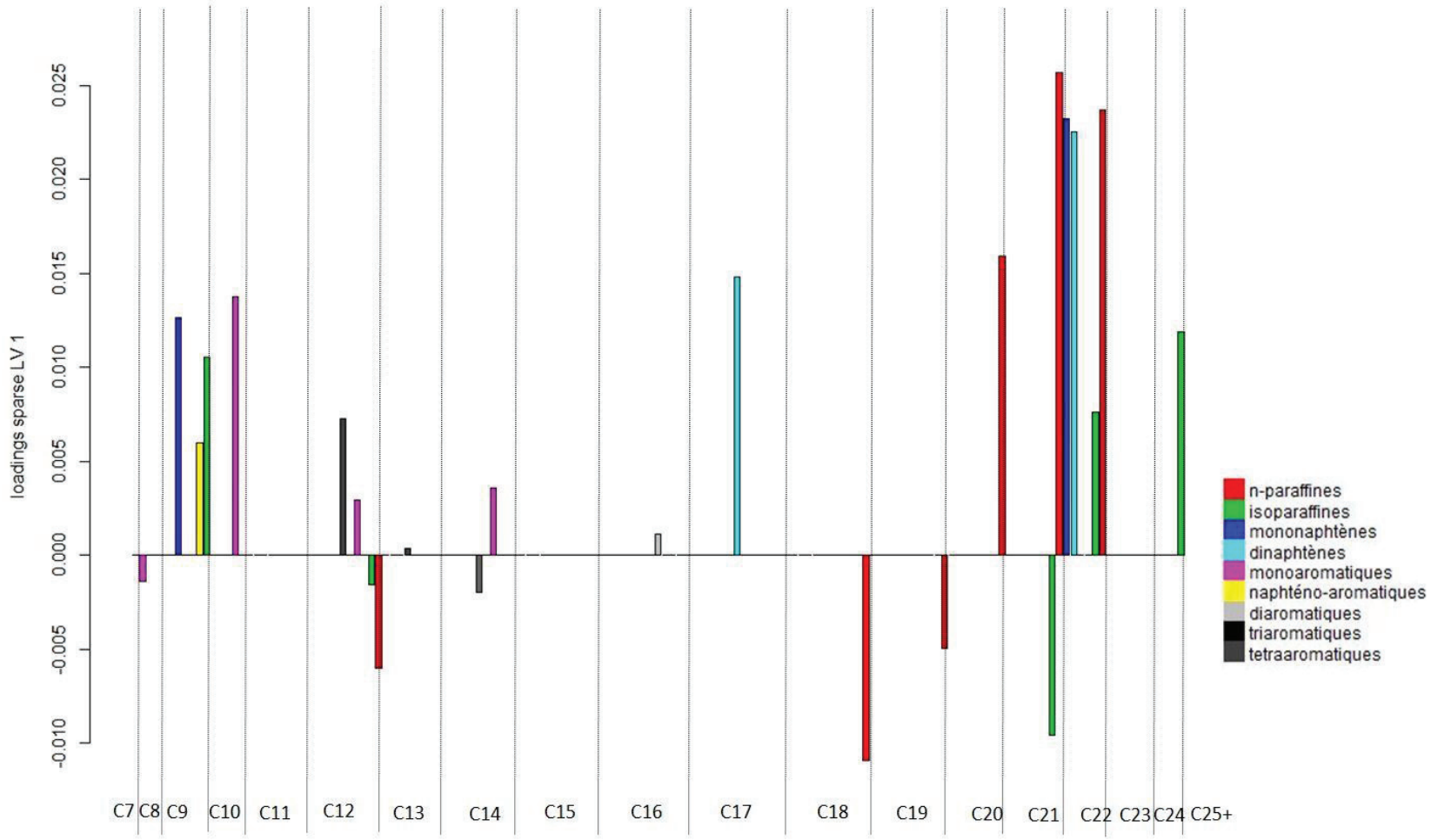


Figure 47 : ***Loadings*** **des variables chromatographiques sur la composante** ***sparse*** **LV1**

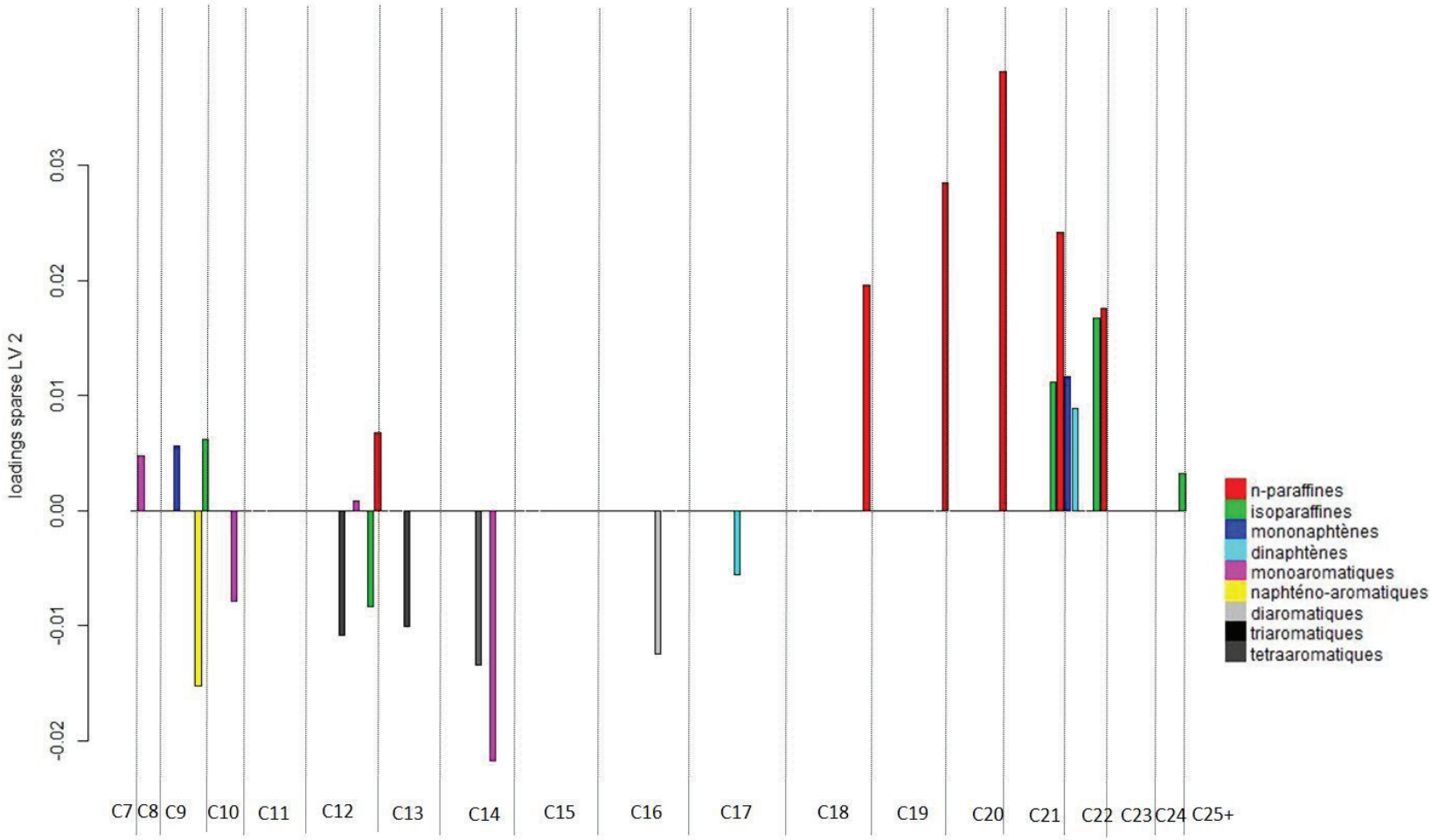


Figure 48 : *Loadings* des variables chromatographiques sur la composante *sparse* LV2

Les coefficients globaux de la *sparse* PLS sont représentés sur la Figure 49 pour chaque composé retenu. On note premièrement une contribution positive importante des composés n-paraffiniques qui ont entre 18 et 22 atomes de carbone. On observe également une contribution négative des isoparaffines qui ont entre 21 et 22 atomes de carbones. Ces deux résultats sont en adéquation avec les observations recensées dans la littérature. La contribution positive des n-paraffines de longues chaînes confirme que le processus de cristallisation de ce type de composé est un facteur essentiel de la variation du point de trouble. La contribution négative des isoparaffines de longueur de chaîne équivalente traduit la capacité de ces molécules à ralentir le processus de cristallisation des n-paraffines. On remarque par ailleurs une contribution positive des composés polyaromatiques à 12 atomes de carbones et de dinaphtènes de 17 et 22 atomes de carbone. Ces dernières observations sont cependant difficiles à interpréter physiquement.

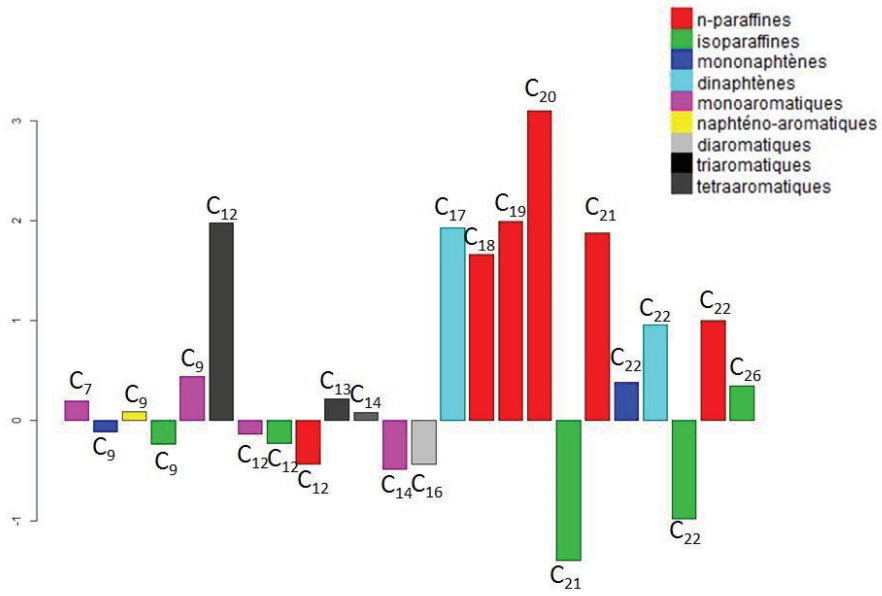


Figure 49 : Coefficients globaux non nuls de la *sparse* PLS pour la modélisation du PT par *sparse* PLS ; orientation des barres verticales : vers le haut ➔ contribution positive ; vers le bas ➔ contribution négative

Globalement, ces résultats sont en adéquation avec ceux qui ont été obtenus par l'approche empirique. On retrouve en effet l'influence des n-paraffines C_{19} à C_{22}, ce même que celle des isoparaffines avec toutefois une précision sur le type d'espèce (C_{21} et C_{22}). Ces résultats sont très cohérents puisqu'ils sous-entendent que ralentissement du processus de cristallisation des n-paraffines serait en grande partie dû à des isoparaffines de taille équivalente à celles des n-paraffines concernées.

L'approche multivariée proposée ci-dessus permet donc d'obtenir des informations très proches de l'approche empirique avec l'avantage de d'une mise en œuvre plus rapide. Dans la suite de nos travaux, seule l'analyse multivariée sera présentée.

6.2 Etude du VI de la coupe huile

Une analyse multivariée a donc été effectuée pour exploiter les données issues d'analyse GC×GC des 31 échantillons d'huile qui ont été collectés (paragraphe 5.2.2) dans le but d'expliquer le VI. Les résultats de cette étude résultats de cette étude sont représentés ci-dessous.

6.2.1 Visualisation de la base de données par ACP

Pour visualiser l'ensemble des échantillons d'huile analysés par GC×GC, une ACP a été réalisée. Les scores des échantillons sur (PC1,PC2) qui représentent 62% de la variance expliquée sont illustrés sur la Figure 50. Les vecteurs de *loadings* associés à chaque composante sont donnés sur la Figure 51 pour PC1 et sur la Figure 52 pour PC2. On observe deux variabilités indépendantes (Figure 50) : une variabilité en majorité des échantillons qui proviennent de charge de type VGO_HDT K suivant PC1 ; une évolution suivant l'axe PC2 qui concerne majoritairement les échantillons issus de la charge VGO _HDT I. Pour l'axe PC1, il apparait clairement que les échantillons de faibles VI sont riches en composés naphténiques et aromatiques. *A contrario*, les huiles de forts VI contiennent une proportion relativement importante de n-paraffines et d'isoparaffines. La nuance « relative » est mentionnée ici pour les n-paraffines en particulier dont le déparaffinage préalable à l'obtention de la coupe huile réduit fortement la teneur. Cette observation est en accord avec les données de la littérature (voir Tableau 3 – paragraphe 2.3.1 et Figure 6 – paragraphe 2.3.4.1) montrant que plus la teneur en saturés croît, plus le VI est élevé. L'observation des *loadings* sur la composante PC2 (Figure 52) confirme ces tendances. Comme pour la coupe gazole, ces divergences de comportement traduisent une différence de composition moléculaire des échantillons suivant leur charge d'origine. Cette observation est par ailleurs cohérente puisque lors d'une même expérimentation, la coupe gazole et la coupe huile sont issue d'un même effluent.

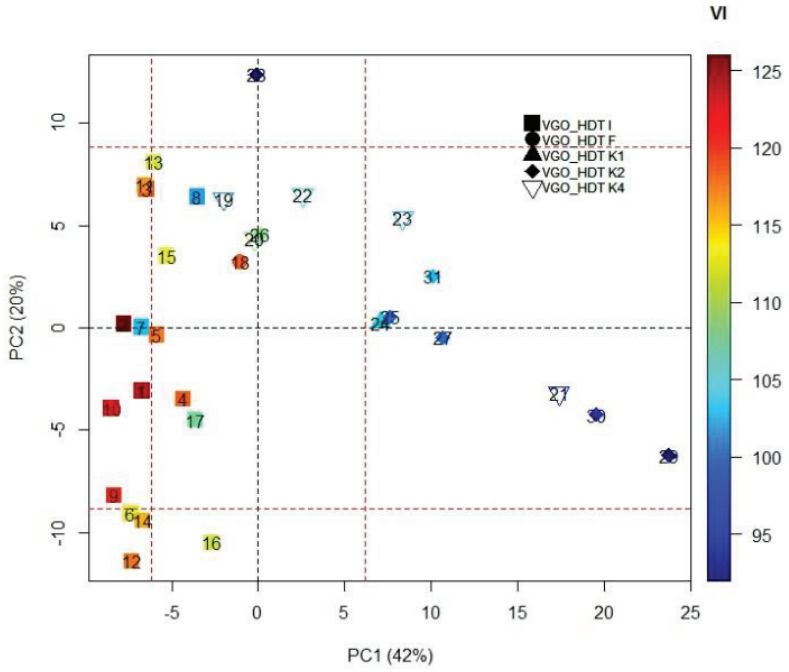


Figure 50 : Scores des échantillons d'huile caractérisés par GC×GC sur les composantes (PC1, PC2)

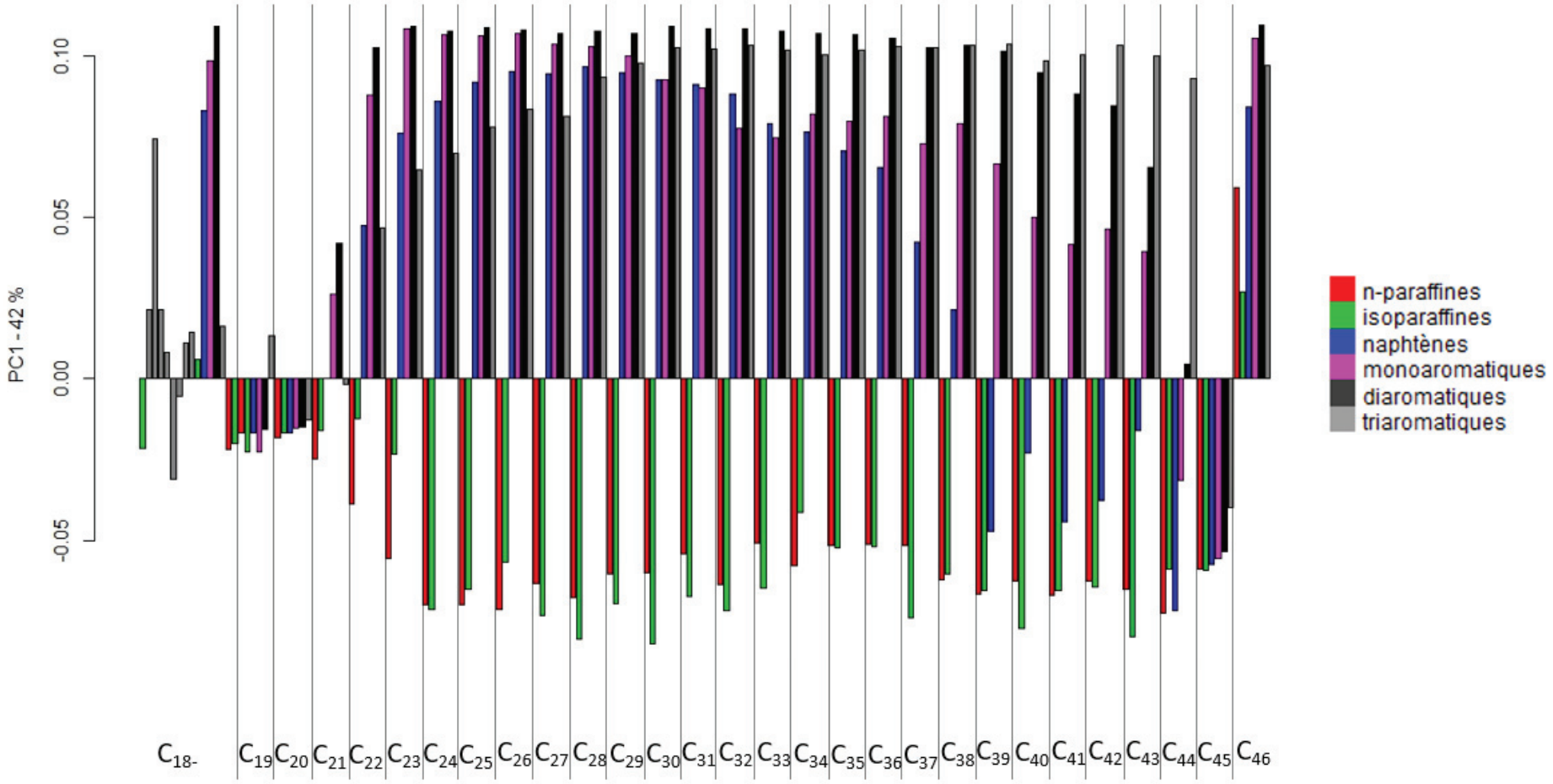


Figure 51 : *Loadings* des variables chromatographiques sur la composante PC1

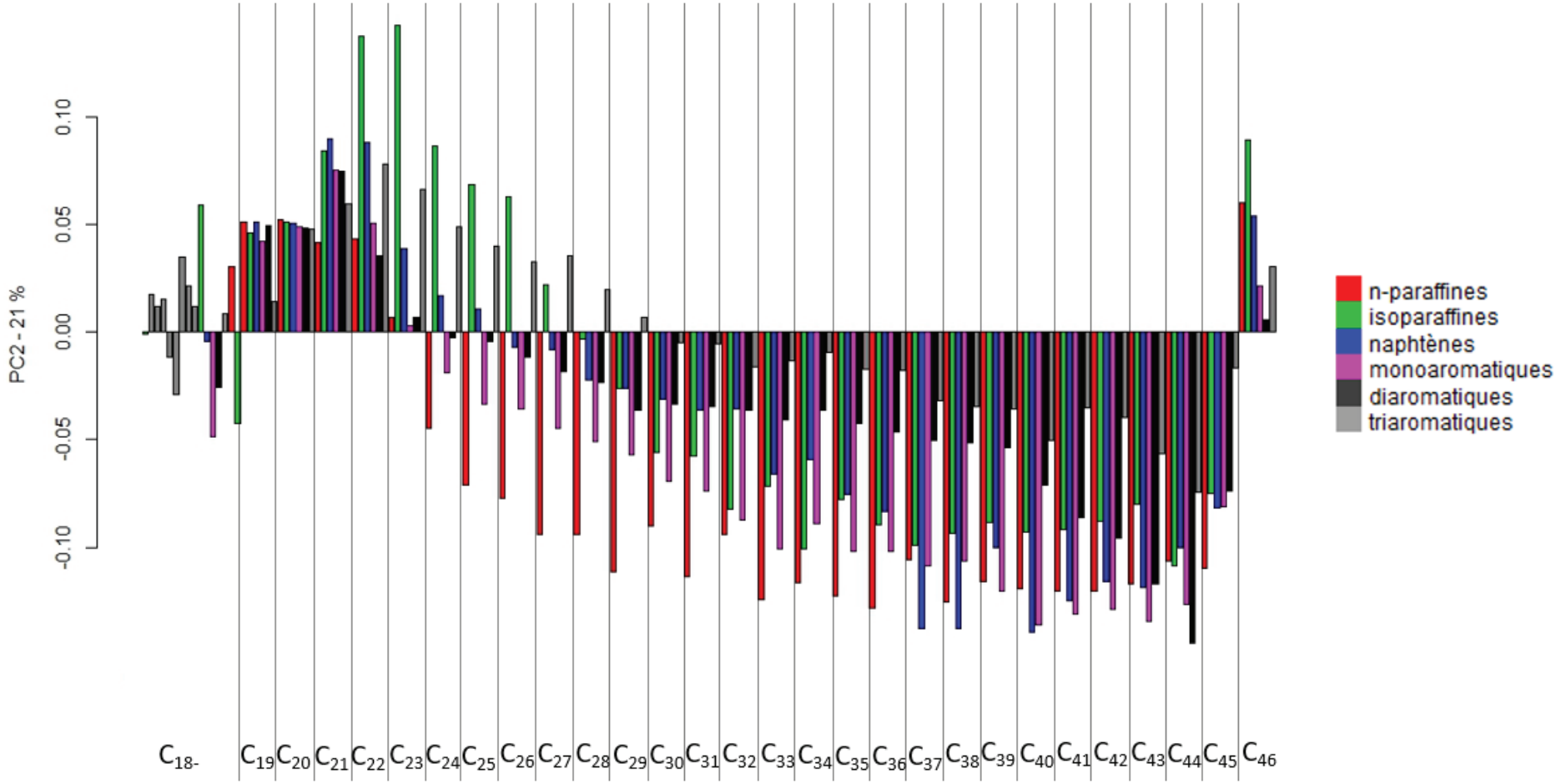


Figure 52 : *Loadings* des variables chromatographiques sur la composante PC2

6.2.2 Prédiction du VI par régression PLS appliquée aux données GC×GC de la coupe huile

Un modèle de régression PLS a été développé afin de vérifier la capacité prédictive de la méthode. Contrairement au cas des gazoles, nous ne disposions pas dans notre base de données d'une réserve d'échantillons. Nous avons donc sélectionné aléatoirement 5 échantillons parmi les 31 dont nous disposions pour constituer une base de test. L'évolution de la RMSECV en fonction du nombre de variables latentes est donnée sur la Figure 53. On observe une valeur minimale de 6,2 pour un nombre de variables latentes égale à 1. Cette valeur de RMSECV est relativement élevée, du fait de la mauvaise prédiction de six échantillons (2, 7, 8, 18, 19, et 28). Ces échantillons n'ont toutefois pas été retirés de la base de données car les analyses réalisées ne présentent pas d'anomalies particulières. De plus, les valeurs de VI de référence ont également été vérifiées.

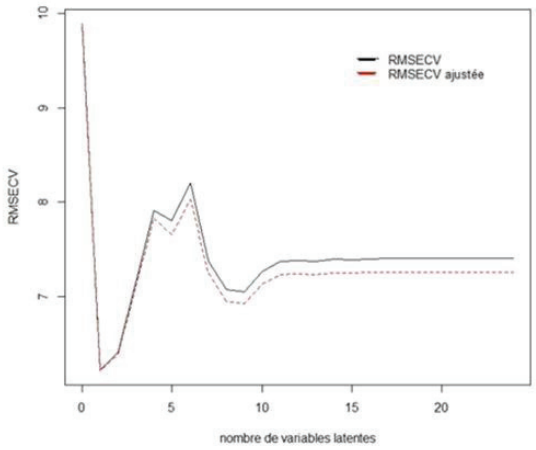


Figure 53 : Evolution de la RMSECV pour l'optimisation du nombre de variables latentes

Le modèle a été évalué sur la base d'apprentissage, puis sur la base de test. Les graphes de parité du modèle pour chacune des bases sont représentés sur la Figure 54 et les statistiques de performance obtenues sont précisées dans le Tableau 26. Elles confirment la difficulté du modèle à prédire certains échantillons. En effet, seulement 35% des échantillons d'apprentissage sont prédits avec une erreur inférieure à l'IC de référence (2,8°C). Ce pourcentage est à peu près équivalent (2 / 5, soit 40%) sur la base de test. Globalement, les performances des modèles obtenus sont clairement insuffisantes. La suite de l'étude et donc l'analyse du modèle *sparse* PLS devra également fournir une explication à cette observation.

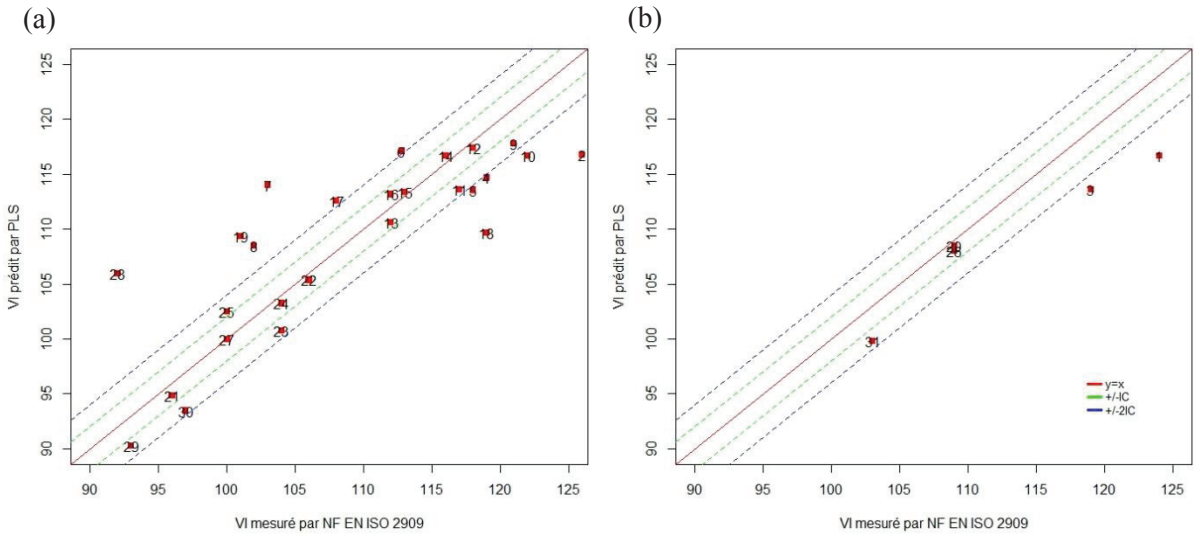


Figure 54 : Graphes de parité du modèle de prédiction du VI par régression PLS appliquée aux données GC×GC ; a) sur la base d'apprentissage ; b) sur la base de test

Tableau 26 : Statistiques du modèle de prédiction du VI de la coupe huile par régression PLS appliquée aux données GC×GC

Base d'évaluation	RMSECV	RMSE*	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)	IC méthode NF ISO 2909
Apprentissage	6,2	5,5	35	58	2
Test	-	2,87	40	60	

* → RMSEC pour la base d'apprentissage ; RMSEP pour la base de test

6.2.3 Modélisation du VI par *sparse* PLS appiquée aux données GC×GC de la coupe huile

Bien que le modèle de régression PLS ne soit pas satisfaisant, une *sparse* PLS a été réalisée sur l'ensemble des échantillons d'huile analysés par GC×GC.

6.2.3.1 Développement du modèle sparse PLS

L'évolution de la RMSECV en fonction du coefficient de seuillage pour un modèle à 1 variable latente est représentée sur la Figure 55. La RMSECV est minimale pour $\eta = 0{,}88$.

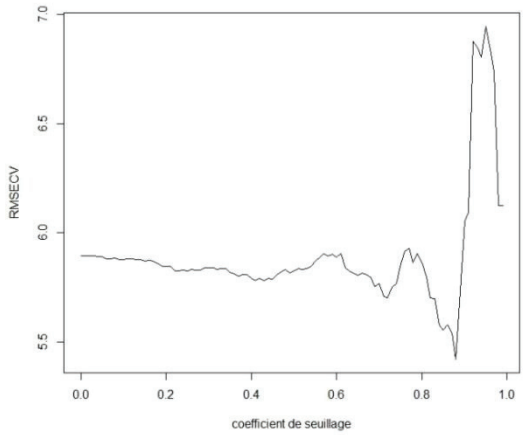


Figure 55 : Evolution de la RMSECV en fonction du coefficient de seuillage pour 1 variable latente

Les graphes de parité des modèles PLS et *sparse* PLS sur la base d'apprentissage sont illustrés sur la Figure 56. Les statistiques correspondantes sont précisées dans le Tableau 27. On note que la valeur de RMSEC est plus élevée pour le modèle PLS que pour le modèle *sparse* PLS (5,31 contre 4,35), de même que le pourcentage de points prédits avec une erreur inférieure à l'IC (55% pour la *sparse* PLS contre 35% pour la PLS). Les graphes de parité montrent que cette différence est notamment liée aux échantillons 1 et 2 qui sont mieux prédits par le modèle parcimonieux (Figure 56b) que par le modèle classique (Figure 56a). Ce constat renvoie aux origines de l'utilisation de la *sparse* PLS. En effet, cette technique a été introduite pour améliorer la robustesse de la PLS classique, trop sensible à la présence de points singuliers [55]. Les échantillons 7, 13, 10, 18, 19, et 28 sont quant à eux toujours mal prédits. L'hypothèse la plus probable est que les données GC×GC des huiles ne permettent pas d'expliquer entièrement la variation du VI. L'absence de données structurales des molécules isoparaffiniques peut constituer une limite à l'utilisation de cette technique dans ce cas précis.

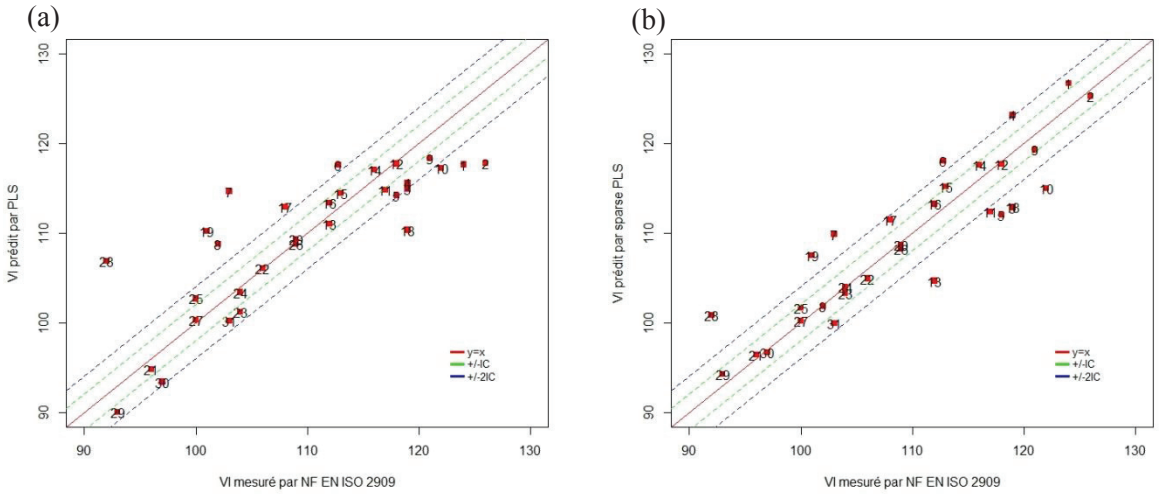


Figure 56 : Graphes de parité sur la base d'apprentissage ; a) pour le modèle PLS à 1 variable latente ; b) pour le modèle *sparse* PLS à une variable latente et $\eta = 0,88$

Tableau 27 : Statistiques du modèle de prédiction du VI de la coupe huile par régression PLS et *sparse* PLS appliquée aux données GC×GC sur la base d'apprentissage

Modèle	Nbre variables latentes	Coefficient de seuillage	RMSECV	RMSEC	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
PLS	1	-	5,9	5,3	35	64
sparse PLS	1	0,88	5,4	4,3	55	74

6.2.3.2 Interprétation de la sparse PLS

Le modèle *sparse* PLS permet d'observer des tendances globales intéressantes. Les coefficients globaux du modèle sont illustrés sur la Figure 57. On retrouve des résultats mentionnés dans l'étude bibliographique qui se traduisent par une opposition nette entre deux types de composés : d'une part les isoparaffines C_{28} à C_{43} qui tendent à augmenter le VI [73,170] et d'autre part les composés aromatiques C_{22} à C_{26} qui ont un effet inverse [73,171]. L'impact négatif des aromatiques sur le VI avait déjà été évoqué par plusieurs auteurs [73,171]. Les n-paraffines étant peu présentes dans la coupe huile en raison du déparaffinage, elles ont peu d'impact.

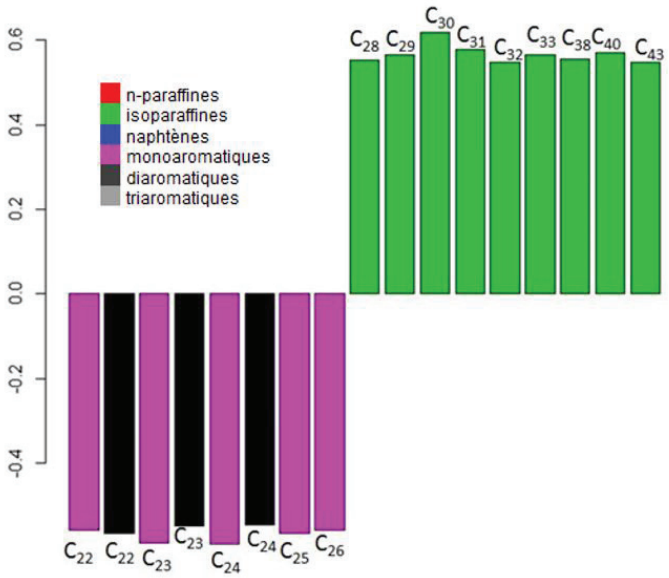


Figure 57 : Coefficients globaux non nuls des variables chromatographiques pour la modélisation du VI par *sparse* PLS

6.3 Conclusion

L'exploitation des données issues de la GC×GC des coupes pétrolières (gazole et huile) a permis de mettre en évidence l'influence prépondérante de certains composés sur les propriétés d'intérêt. La méthodologie suivante a été mise en place : 1) visualisation de données pour dégager des tendances globales ; 2) analyse multivariée par PLS pour vérifier l'impact des paramètres d'entrée sur une grandeur d'intérêt (PT ou VI dans notre cas) ; 3) analyse multivariée par *sparse* PLS pour déterminer les paramètres les plus influents sur la propriété d'intérêt, ceci permet de mieux comprendre cette propriété.

Dans le cas des gazoles, cette étude a permis de confirmer l'impact des composés n-paraffiniques et isoparaffiniques sur la variation du PT (déjà évoqué dans la littérature). Les n-paraffines les plus lourdes sont celles qui ont tendance à cristalliser en premier. Les isoparaffines qui ont une structure très proche de ces n-paraffines ralentissent leur processus de cristallisation. Le modèle obtenu est de bonne qualité : la quasi-totalité des échantillons mis en jeu sont prédits dans l'intervalle de confiance de la mesure de référence.

Dans le cas des huiles, les modèles obtenus sont d'une qualité moindre. Ceci montre que la GC×GC ne contient pas l'information pertinente pour expliquer l'évolution du VI. Seules les tendances globales de la littérature sont retrouvées : l'impact prépondérant des isoparaffines et des aromatiques sur le VI.

Ces études vont être complétées par application sur une autre technique analytique : la RMN du ^{13}C.

Chapitre 7. Compréhension moléculaire des propriétés produits par RMN du ^{13}C

Les études menées au chapitre précédent sur les données issues d'analyse GC×GC de la coupe gazole et de la coupe huile a permis de montrer l'efficacité de la méthodologie d'exploitation proposée dans cette étude.

Il est apparu que la caractérisation d'échantillons par GC×GC était pertinente pour prédire le PT de la coupe gazole, mais pas dans le cas du VI. Pour compléter cette étude, la même méthodologie a été appliquée sur une seconde technique de caractérisation d'échantillons : la RMN du ^{13}C. Les cas d'application sont les mêmes que précédemment : VI de la coupe huile et PT de la coupe gazole. Ces résultats seront utilisés pour mieux comprendre le comportement de deux catalyseurs d'HCK.

7.1 Etude du VI de la coupe huile

L'objet de ce paragraphe est de présenter, commenter et discuter les résultats de l'exploitation par méthode multivariée des données issues de l'analyse RMN du ^{13}C de l'ensemble des échantillons d'huile collectés pour la compréhension moléculaire du VI. La méthodologie utilisée est la même que celle qui a été appliquée au données GC×GC.

7.1.1 Visualisation des échantillons par classification ascendante hiérarchique

Pour cette étude, deux échantillons d'huile ont été retirés de la base de données en raison d'un décalage significatif malgré le prétraitement (paragraphe 3.2.2.4). Le reste de la base de données (29 échantillons) est illustré sur le dendrogramme de la Figure 58. Il a été obtenu en réalisant une CAH des spectres ^{13}C RMN des échantillons d'huile après prétraitement. Les numéros et les valeurs de VI des échantillons sont spécifiés.

On note la présence de trois principaux *clusters* caractérisés par les nœuds (a), (b) et (c) (Figure 58). Les échantillons des clusters caractérisés par les nœuds (a) et (b) sont en majeure partie issus de charges filles VGO_HDT K. Dans le cas du nœud (a) on retrouve majoritairement les échantillons provenant de VGO_HDT I. Cela traduit encore une fois une variabilité des spectres en fonction de la charge d'origine.

On remarque par ailleurs que les deux échantillons dont les spectres ^{13}C RMN sont les plus proches (19 et 28 tout en bas du dendrogramme) ont des VI équivalents (96 et 97 respectivement). Ce constat peut être étendu à l'ensemble de la base de données puisque dans la majeure partie des cas, les regroupements de plus bas niveaux concernent des échantillons de VI proches. On peut donc légitimement penser que certaines zones des spectres ^{13}C RMN sont liées à la variation du VI.

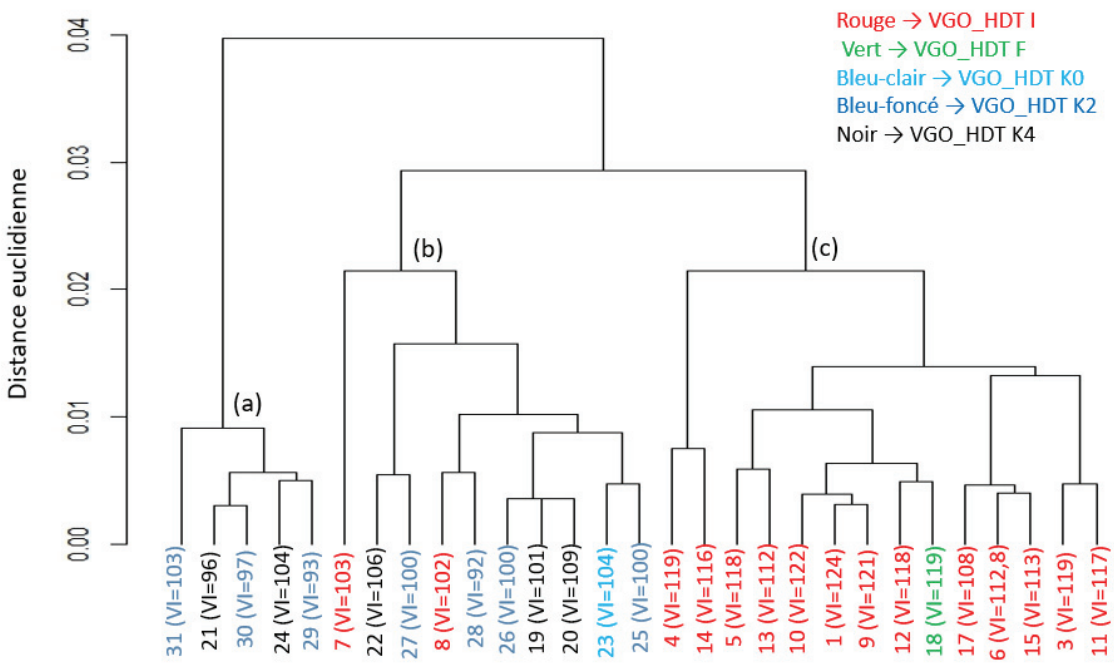


Figure 58 : Dendrogramme obtenu par classification ascendante hiérarchique des spectres ^{13}C RMN des échantillons d'huile

7.1.2 Prédiction du VI par régression PLS appliquée aux données de ^{13}C RMN de la coupe huile

Une régression PLS a été effectuée sur les données issues des analyses RMN du ^{13}C des échantillons d'huile (paragraphe 2.5). La base de données a été divisée en deux : une base d'apprentissage de 26 échantillons et une base de test de 5 échantillons. La sélection des échantillons de test a été faite de manière aléatoire. La Figure 59 représente l'évolution de l'erreur quadratique moyenne de validation croisée (RMSECV) en fonction du nombre de variables latentes. Cette RMSECV est minimale pour un nombre de variables latentes égal à 2.

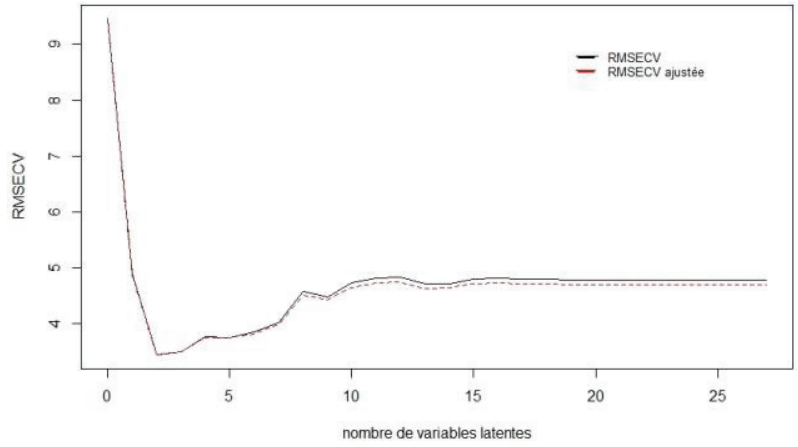


Figure 59 : Evolution de la RMSECV en fonction du nombre de variables latentes

Les graphes de parité du modèle PLS obtenu sont représentés sur la Figure 60 aussi bien sur la base d'apprentissage que sur la base de test. Les performances du modèle sont précisées dans le Tableau 28. Sur les données d'apprentissage, on note que le VI de l'échantillon 26 est particulièrement mal prédit (Figure 60a). Pour le reste de la base, la qualité de la prédiction est satisfaisante puisqu'une RMSEP de 2,6 a été obtenue et plus de la moitié des points (54%) des échantillons d'apprentissage sont prédits avec une erreur inférieure à l'IC de la mesure de référence. Sur la base de test, 4 échantillons sur 5 sont relativement bien prédits. La valeur de la RMSEP est de 3,2 et est principalement due à l'échantillon mal prédit. Le modèle de régression PLS appliquée aux données ^{13}C RMN des huiles permet globalement d'estimer le VI de manière satisfaisante.

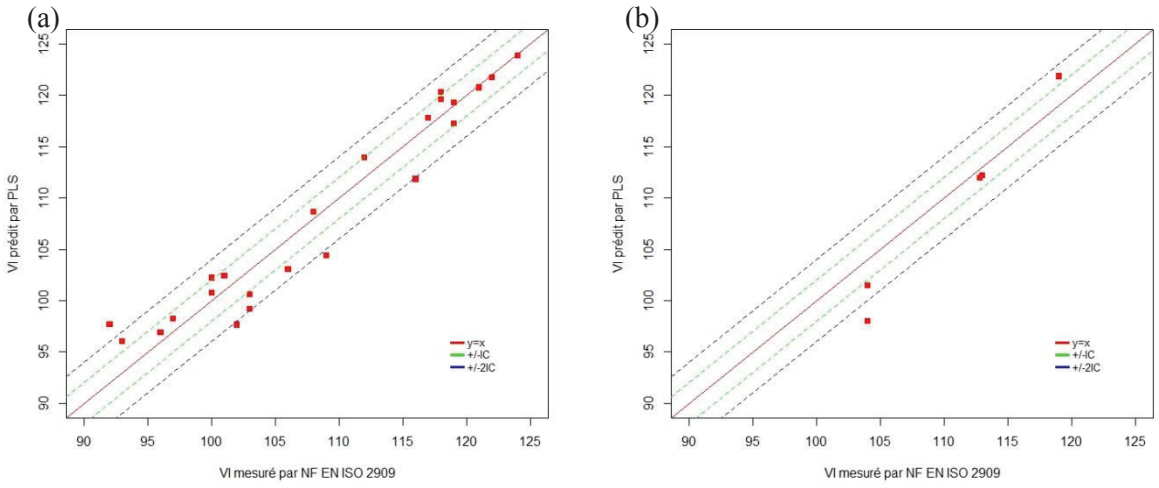


Figure 60 : Graphes de parité du modèle de régression PLS appliquée aux données ^{13}C RMN des échantillons d'huile pour la prédiction du VI ; a) sur la base d'apprentissage ; b) sur la base de test

Tableau 28 : Statistiques du modèle de prédiction du VI de la coupe huile par régression PLS appliquée aux données ^{13}C RMN

Base d'évaluation	RMSECV	RMSE*	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)	IC méthode NF ISO 2909 (°C)
Apprentissage	3,2	2,6	54	83	2
Test		3,2	40	80	

* → RMSEC pour la base d'apprentissage ; RMSEP pour la base de test

7.1.1 Modélisation du VI par régression *sparse* PLS appliquée aux données ^{13}C RMN de la coupe huile

Un modèle de régression PLS et un modèle de *sparse* PLS ont été développé à partir de l'ensemble des données ^{13}C RMN des échantillons d'huile. Le nombre de variables latentes optimisé est toujours égal à 2 (malgré la réintroduction des 5 spectres précédemment utilisés comme base de test). L'évolution de la RMSECV en fonction de η (pour $0 \leq \eta < 1$) et pour un nombre de variables latentes égal à 2 est représentée sur la Figure 61. La RMSECV minimale est obtenue pour $\eta = 0{,}75$.

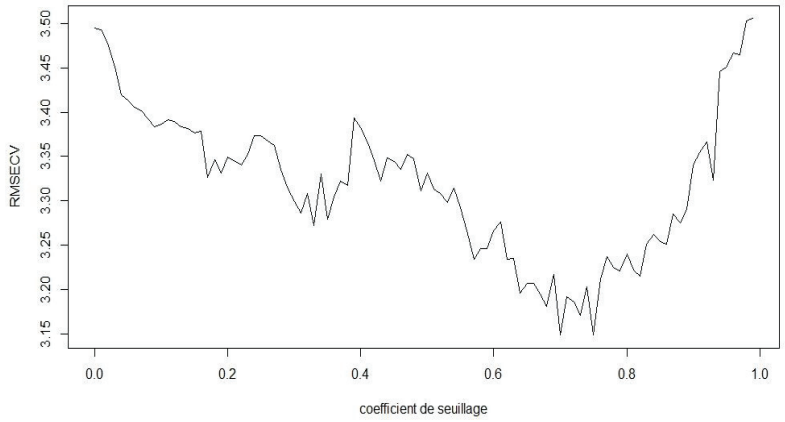


Figure 61 : Evolution de l'erreur quadratique moyenne de validation croisée en fonction du coefficient de seuillage

Les graphes de parité des modèles PLS et *sparse* PLS sur les données d'apprentissage sont illustrés sur la Figure 62. Ces graphes de parité semblent clairement équivalents. Ce constat est appuyé par les statistiques correspondantes qui sont précisées dans le Tableau 29. En effet, la RMSEC et les pourcentages de points prédits dans l'IC de la mesure sont identiques (2,3 et 69% respectivement).

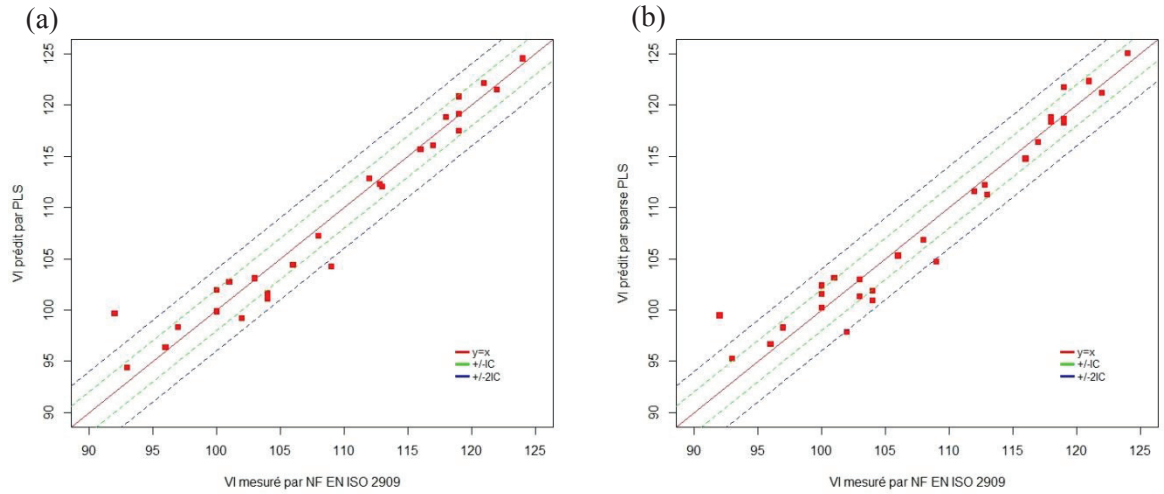


Figure 62 : Graphes de parité sur la base d'apprentissage pour la modélisation du VI à partir des données ^{13}C RMN ; a) modèle PLS ; b) modèle *sparse* PLS

Tableau 29 : Statistiques du modèle de prédiction du VI de la coupe huile par régression PLS et *sparse* PLS appliquées aux données ^{13}C RMN sur la base d'apprentisssage

Modèle	Nombre de variables latentes	Coefficient de seuillage	RMSECV	RMSEC	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
PLS	2	-	3,5	2,3	69	89
sparse PLS	2	0,75	3,1	2,3	69	89

7.1.1.1 Interprétation de la sparse PLS

Les *loadings* associés aux 2 composantes parcimonieuses ainsi que les coefficients globaux du modèle *sparse* PLS sont représentés sur la Figure 63. La composante *sparse* LV1, dont les *loadings* sont illustrés sur la Figure 63a est la plus fortement corrélée au VI (coefficient de corrélation de 0,87). On note que les zones les plus influentes sont localisées autour de 14,1, 29,9 et 30,3 ppm. Ces zones d'influence sont identiques pour la composante *sparse* LV2 (Figure 63b). Les coefficients globaux estimés par le modèle sont représentés sur la Figure 63c. On observe :

- une influence négative des variables situées autour de 14,2 ppm (zone caractéristique de carbones de type CH_3 au bout d'une chaîne alkyle de plus de six carbones, Figure 65)
- une influence positive des variables situées autour de 29,9 ppm (zone caractéristique de carbone de type CH_2 au milieu d'une chaîne droite, Figure 65)
- une influence positive des variables situées autour de 30,3 ppm (zone caractéristique de carbone de type CH_2 en position γ par rapport à un branchement méthyle, Figure 65)

On retrouve ainsi les résultats des études précédentes [73] : l'impact positif des carbones n-paraffiniques et des carbone CH_2 en position γ ou β par rapport à un branchement méthyle. Le pic à 14,2 ppm n'avait pas été étudié par Verdier (impact négatif de chaînes méthyl sur le VI). La base de données étudiées ici est différente car elle contient exclusivement des échantillons d'huile produits par HCK (qui ont une très faible proportion de n-paraffines). Des contributions isolées sont observables à proximité des pics d'influence (traits fins, Figure 63). Cela montre que l'alignement des spectres n'est pas optimal.

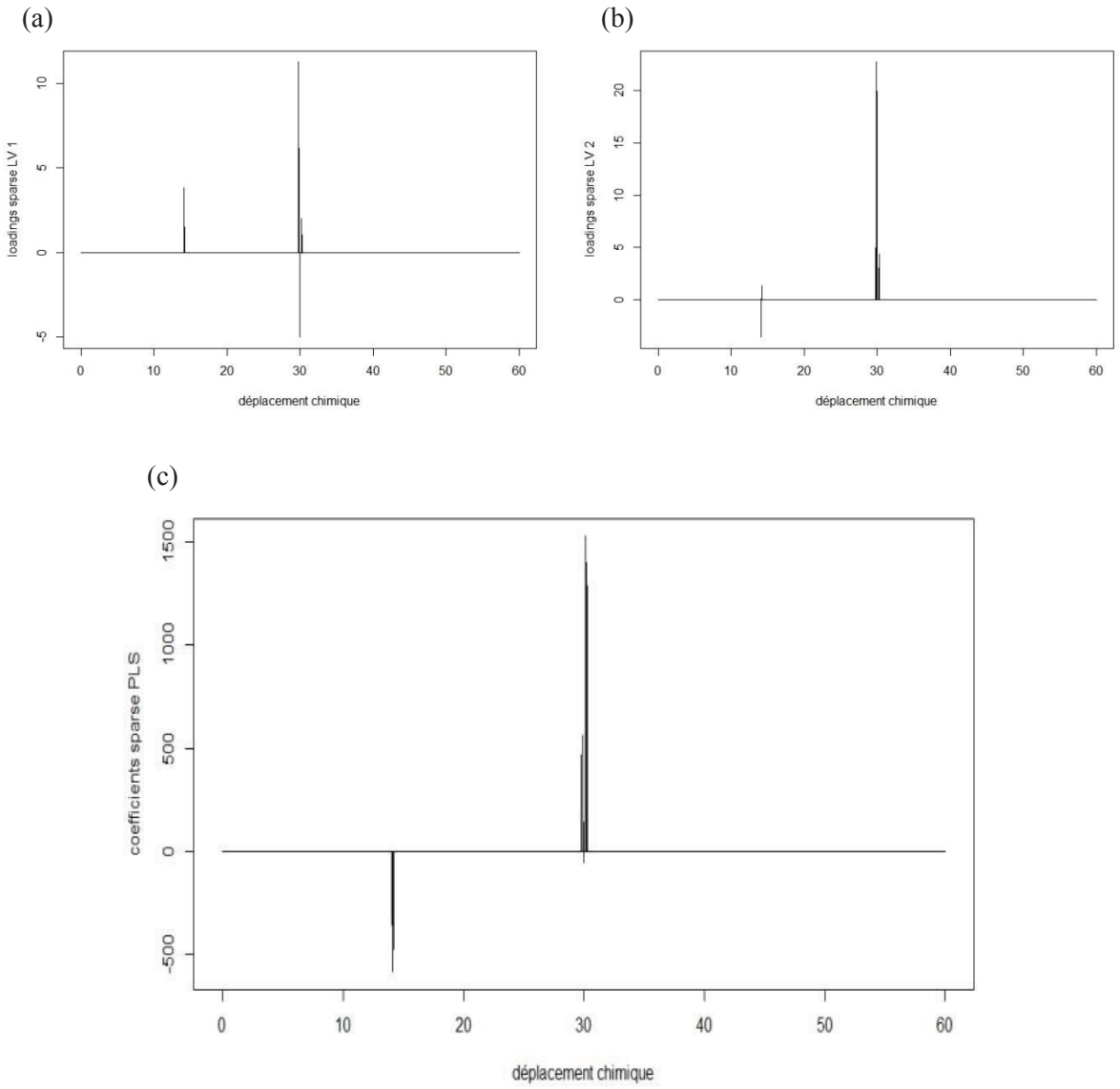


Figure 63 : Modèle *sparse* PLS ; a) *Loadings* sur *sparse* LV 1 ; b) *Loadings* sur *sparse* LV 2 ; c) coefficients globaux de la *sparse* PLS

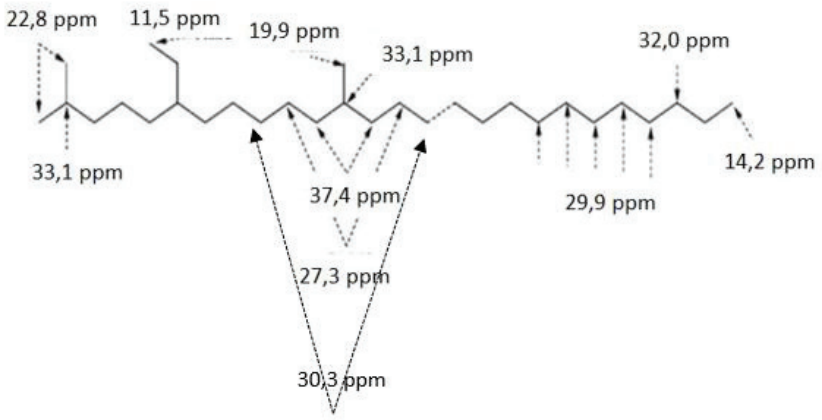


Figure 64 : Structures moléculaires identifiables par ^{13}C RMN et déplacements chimiques correspondants [73]

La méthodologie proposée dans cette thèse a donc abouti à des résultats en adéquation avec les travaux de Verdier *et al.* [73] qui sont une référence dans ce domaine. Ceci permet une fois de plus de valider cette méthodologie. L'approche introduite ici est d'autant plus intéressante qu'elle permet d'obtenir des résultats similaires sans être contraint au préalable d'intégrer tous les pics caractéristiques du spectre comme c'est le cas dans l'étude de Verdier *et al.*[73]. Cette étude montre également que la RMN du ^{13}C permet d'apporter des informations complémentaires à celles de la GC×GC. Notamment, les échantillons mal prédits à partir des données GC×GC (paragraphe 6.2.3) sont en revanche bien prédits à partir des données ^{13}C RMN. Les informations concernant la structure des molécules semblent donc prépondérantes pour expliquer le VI.

7.2 Etude du PT de la coupe gazole

Une étude de compréhension moléculaire du PT de la coupe gazole à partir de spectres ^{13}C RMN a donc été effectuée. Dans ce paragraphe, nous présentons et discutons les résultats de cette étude. La méthodologie utilisée est la même que dans le cas des huiles.

7.2.1 Analyse exploratoire des échantillons par CAH

De même que pour l'étude du VI de la coupe huile, une CAH a été réalisée sur les spectres ^{13}C RMN des échantillons de gazole. Rappelons que 6 spectres ont été retirés en raison d'un décalage trop important malgré le prétraitement effectué (paragraphe 3.2.2.3). La base de données est donc constituée ici de 34 échantillons. Le dendrogramme obtenu est représenté sur la Figure 55. Les numéros et les valeurs de PT mesurées des différents échantillons sont également spécifiés. On observe dans ce cas la présence de deux *clusters* caractérisés par les nœuds (a) et (b) (Figure 55). Comme dans le cas des huiles ces *clusters* traduisent des différences de spectres entre les échantillons suivant la charge d'origine. On peut de plus noter que globalement, les échantillons qui ont les spectres les plus proches ont un PT voisin

(3 et 5, 11, 17 et 16, ou encore 26 et 35). Les signaux RMN sont donc là encore susceptibles de contenir des informations liées à la variation du PT.

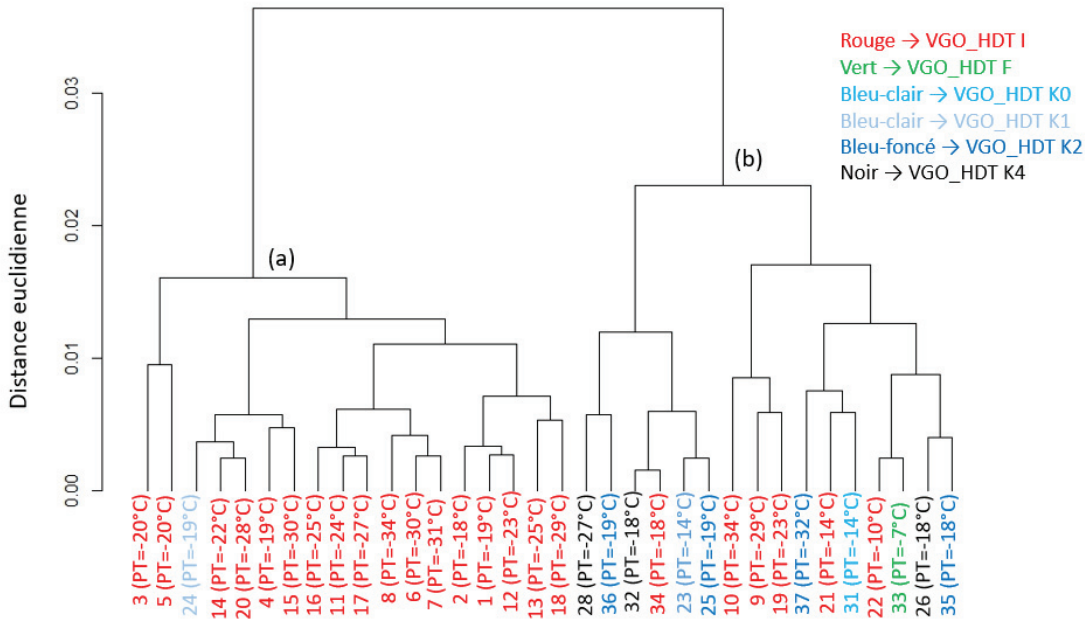


Figure 65 : Dendrogramme obtenu par classification ascendante hiérarchique des spectres ^{13}C RMN des échantillons d'huile

7.2.2 Prédiction du point de trouble par régression PLS appliquée aux données ^{13}C RMN de la coupe gazole

Comme dans le cas des huiles, un modèle de régression PLS a été développé afin de confirmer la pertinence de l'utilisation des données ^{13}C RMN des échantillons de gazole pour modélisation du PT. Parmi ces échantillons, 5 ont été sélectionnés aléatoirement pour former une base de test, le reste constituant la base d'apprentissage. L'évolution de la RMSECV en fonction du nombre de variables latentes est illustrée sur la Figure 66. La valeur minimale est atteinte pour 10 variables latentes.

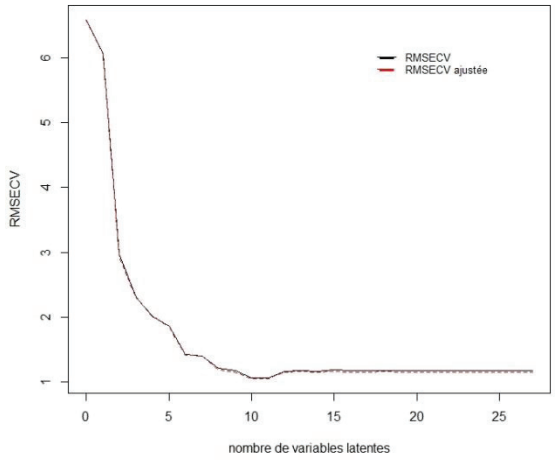


Figure 66 : Evolution de la RMSECV en fonction du nombre de variables latentes pour la prédiction du PT par PLS appliquée aux données ^{13}C RMN de la coupe gazole

Les graphes de parité du modèle PLS sur les bases d'apprentissage et de test sont représentés sur la Figure 67. Les statistiques de performances du modèle sont précisées dans le

Tableau 30. Globalement, la précision du modèle est clairement satisfaisante. En effet, sur la base d'apprentissage, une RMSEP de 0,7°C et 100% des points sont prédits à l'intérieur de l'intervalle de confiance de la mesure de référence. Sur la base de test la RMSEP est de 2,77°C et 3 échantillons sur 5 sont prédits avec une erreur inférieure à l'IC de référence. Ces performances sont clairement satisfaisantes. L'écart entre les valeurs de RMSEP obtenues sur chaque base laisse toutefois penser que le modèle est potentiellement surparamétré du fait d'une quantité encore insuffisante d'échantillons.

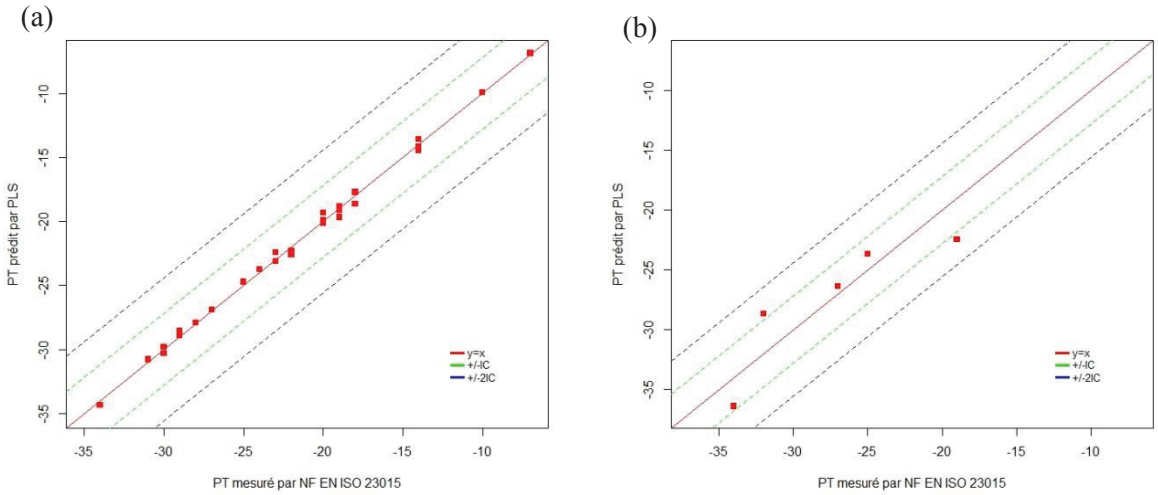


Figure 67 : Graphes de parité du modèle de régression PLS pour la prédiction du PT à partir de données ^{13}C RMN de la coupe gazole ; a) sur la base d'apprentissage ; b) sur la base de test

Tableau 30 : Statistiques du modèle de prédiction du PT de la coupe gazole par régression PLS appliquée aux données ^{13}C RMN

Base d'évaluation	RMSECV (°C)	RMSE* (°C)	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)	IC méthode NF ISO 23015 (°C)
Apprentissage	1,06	0,7	100	100	2,8
Test		2,77	60	100	

* → RMSEC pour la base d'apprentissage ; RMSEP pour la base de test

7.2.3 Modélisation du PT par *sparse* PLS appliquée aux données ^{13}C RMN de la coupe gazole

La qualité satisfaisante du modèle PLS obtenu à partir des données ^{13}C RMN montre qu'elles contiennent des informations susceptibles d'être reliées au PT. Une régression *sparse* PLS a donc été réalisée pour confirmer cette assertion et identifier les zones d'intérêt du spectre.

7.2.3.1 Optimisation des paramètres de la sparse PLS

Une régression PLS a donc à nouveau été effectuée cette fois-ci à partir des 34 échantillons de gazole. L'évolution de la RMSECV en fonction du nombre de variables latentes est donnée sur la Figure 68a. Une valeur de RMSECV minimale a été obtenue pour 8 variables latentes. La courbe d'évolution de la RMSECV en fonction de η (pour $0 \leq \eta < 1$) et pour 8 variables latentes a été tracée sur la Figure 68b. La valeur minimale est obtenue pour $\eta = 0{,}85$.

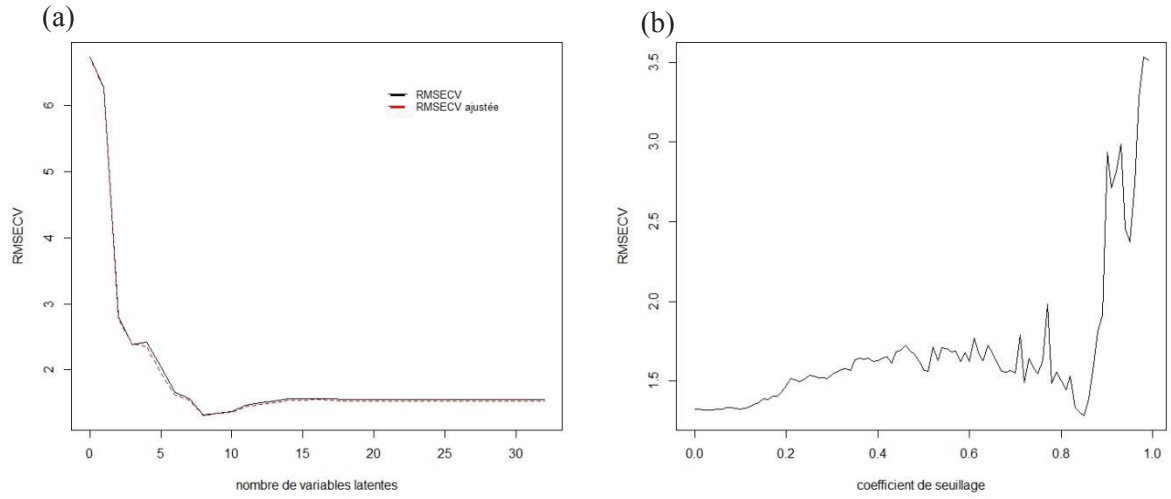


Figure 68 : Modélisation du PT de la coupe gazole ; a) Evolution de la RMSECV en fonction du nombre de variables latentes pour le modèle PLS ; b) Evolution de la RMSECV en fonction du coefficient de seuillage pour un nombre de variables latentes égal à 8

Les performances des modèles PLS et *sparse* PLS estimées sur la base d'apprentissage sont précisées dans le Tableau 31 et les graphes de parité correspondant sont illustrés sur la Figure 69. On peut noter une forte proximité entre les deux modèles. En effet, la RMSEC est de 0,9°C pour le modèle

PLS contre 0,8°C pour le modèle parcimonieux et dans les deux cas 100% des échantillons sont prédits à l'intérieur de l'IC de la mesure de référence. Dans ce cas précis, l'utilisation de la *sparse* PLS parait tout à fait pertinente étant donné que le modèle PLS initial n'est pas dégradé.

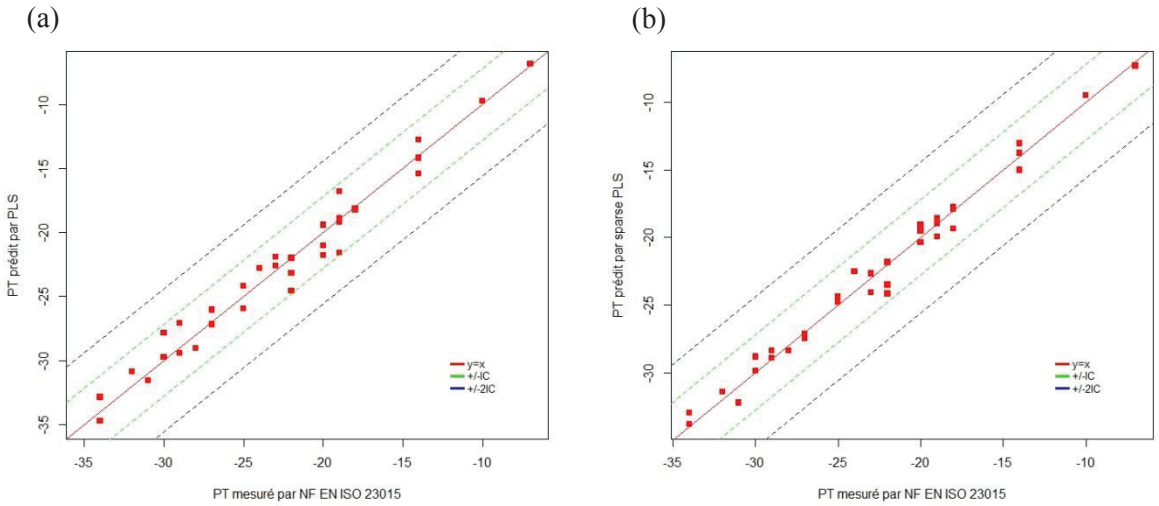


Figure 69 : Graphes de parité sur la base d'apprentissage pour la modélisation du PT à partir des données ^{13}C RMN ; a) modèle PLS ; b) modèle *sparse* PLS

Tableau 31 : Statistiques du modèle de prédiction du PT de la coupe gazole par régression PLS et *sparse* PLS appliquées aux données ^{13}C RMN

Modèle	Nombre de variables latentes	Coefficient de seuillage	RMSECV (°C)	RMSEC (°C)	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
PLS	8	-	1,9	0,9	100	100
sparse PLS	8	0,85	1,3	0,8	100	100

7.2.3.2 Interprétation du modèle sparse PLS

La Figure 70 illustre les vecteurs de *loadings* associés aux deux premières composantes de la *sparse* PLS. La deuxième composante (*spars*e LV2) est la plus corrélée à cette propriété (coefficient de corrélation de 80%). On observe que les zones les plus importantes pour expliquer le PT sont les suivantes (Figure 70b) (voir Figure 65 pour les structures correspondantes) :

- autour du pic à 14,2 ppm (caractéristique d'un carbone de type CH_3)
- autour de 19,9 ppm (caractéristique d'un carbone de type CH_3 branché sur un carbone CH situé à au moins 4 carbones d'un CH_3 terminal)
- autour de 22,8 ppm (caractéristique d'un branchement méthyle en position β par rapport à un méthyle terminal).

- autour de 27,2 ppm (caractéristique d'un méthylène en position β par rapport à un branchement propyl – hexyl)
- autour de 29,9 ppm (caractéristique d'un méthylène en milieu d'une chaîne droite
- autour de 37,4 ppm (caractéristique d'un méthylène en position γ par rapport à un branchement méthyle)

On retrouve les mêmes zones spectrales d'intérêts dans le cas de la composante *sparse* LV1 (Figure 70a).

Les coefficients globaux de la *sparse* PLS qui sont représentés respectivement sur la Figure 70c ont permis d'observer certaines tendances. Les échantillons qui ont une forte proportion de méthylène (CH_2) en milieu d'une chaîne droite (pic à 29,9 ppm) tendent à avoir un PT haut. Or cette caractéristique est particulièrement abondante dans les molécules n-paraffiniques notamment. Toutes les autres structures d'influence qui ont été listées ci-dessus tendent à diminuer le PT et ont en commun de faire chacune référence à la présence de carbones isoparaffiniques (Annexe E). Ces deux constats sont en adéquation avec les conclusions faites à partir de l'étude des propriétés à froid par GC×GC (paragraphe 6.3).

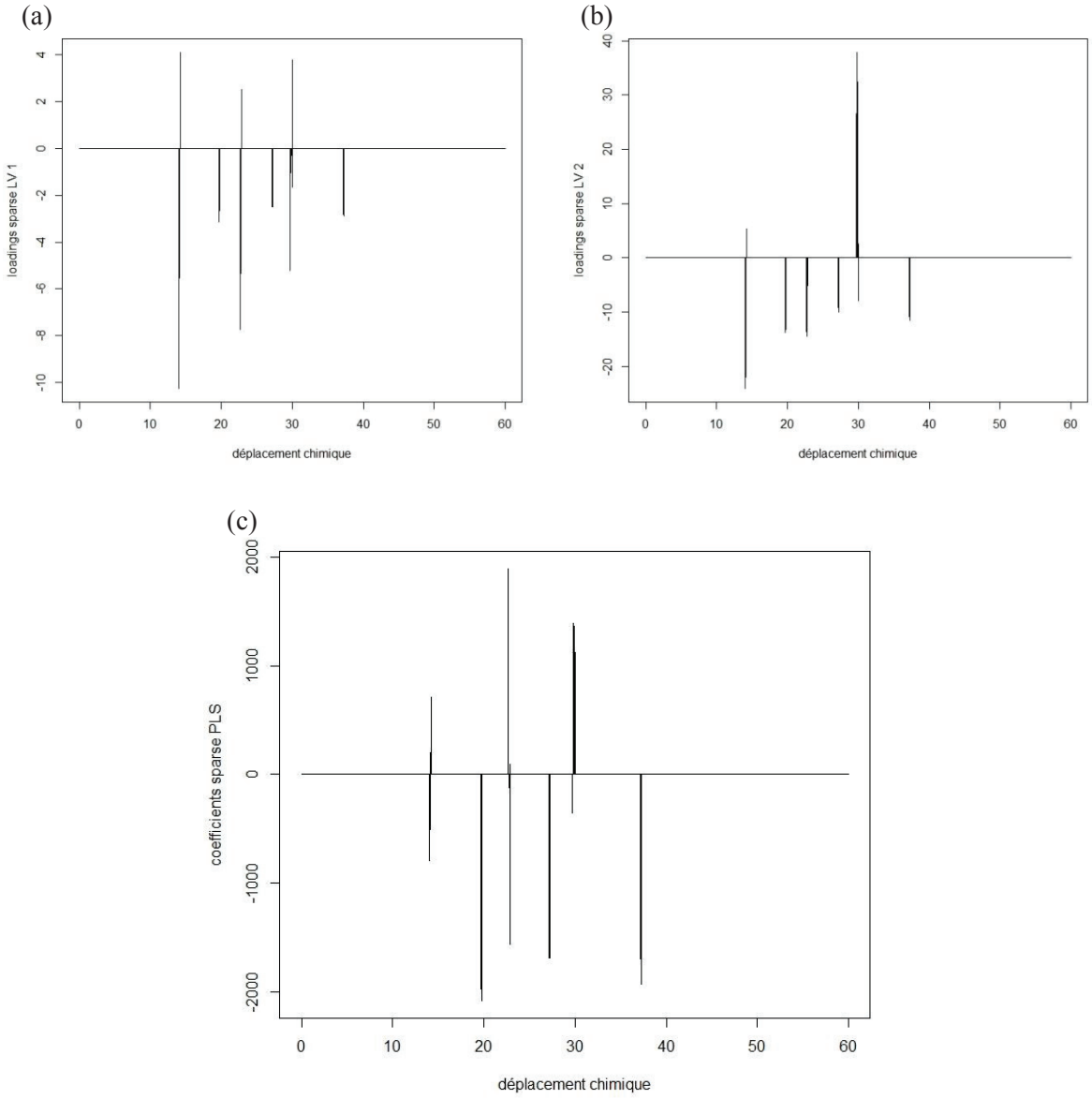


Figure 70 : Modèle *sparse* PLS ; a) *Loadings* sur *sparse* LV 1 ; b) *Loadings* sur *sparse* LV 2 ; c) coefficients globaux de la *sparse* PLS

NB : des alternances de contribution positive / négative autour d'un même déplacement chimique sont observables dans certaines zones sur la Figure 70. Leur présence est essentiellement due à des effets de décalage encore légèrement présents pour certains spectres. Ce qui remet en question la qualité de l'alignement.

L'étude précédente montre donc l'impact positif des chaînes linéaires droites et négatifs des chaînes ramifiées sur le PT. Rappelons que les spécifications européennes requièrent un PT inférieur à -10°C. Pour respecter cette spécification, il est donc suggéré d'avoir des catalyseurs avec un fort pouvoir isomérisant. Ceci sera étudié dans le prochain paragraphe.

7.3 Cas d'application : comparaison de deux catalyseurs

L'étude présentée dans ce paragraphe a été réalisée sur 8 échantillons dont 4 de gazole et 4 d'huile. Dans chaque cas (gazole et huile), les échantillons étudiés sont issus de la même charge DSV prétraitée (VGO_HDT I) et ont été produits dans des conditions opératoires voisines mais en utilisant des catalyseurs différents. Les catalyseurs mis en jeu sont HCK_A, HCK_B ainsi que les deux catalyseurs résultant de leurs empilements (HCK_{B+A} et HCK_{A+B}). Comme nous l'avons illustré au paragraphe 5.2.3, les catalyseurs HCK_A et HCK_B ont sur une molécule modèle n-C_{16} des effets isomérisant et craquant différents (HCK_A est plus craquant que HCK_B). Ceci conduit à des effets opposés sur la qualité des produits, le premier fournissant des huiles de très bonne qualité au détriment de la qualité des gazoles, le second permettant d'améliorer les propriétés à froid des gazoles tout en dégradant la qualité de l'huile. HCK_{B+A} et HCK_{A+B} produisent quant à eux des produits de qualité intermédiaire qui peuvent être intéressants en fonction de la cible visée. L'objectif de cette approche est de mettre en évidence des différences significatives entre les spectres qui puissent éventuellement être reliées au VI dans le cas des huiles et au PT dans le cas des gazoles. La méthodologie utilisée pour cette comparaison des spectres ^{13}C RMN est celle qui a été décrite au paragraphe 3.2.2.3.

7.3.1 Comparaison des spectres d'échantillons d'huile

7.3.1.1 Observation des spectres

La Figure 71 représente les spectres ^{13}C RMN des échantillons produits avec HCK_A et HCK_B uniquement sur la zone des carbones saturés (0 à 60 ppm). On remarque que tous les pics présents dans l'un sont aussi répertoriés dans l'autre. Par contre les intensités relatives de ces pics diffèrent clairement suivant l'échantillon. Le spectre résultant de la différence des deux spectres superposés sur la Figure 71 confirme ces différences entre les intensités (Figure 72). Le choix du catalyseur a donc un impact fort sur les caractéristiques chimiques des effluents même à iso conditions opératoires.

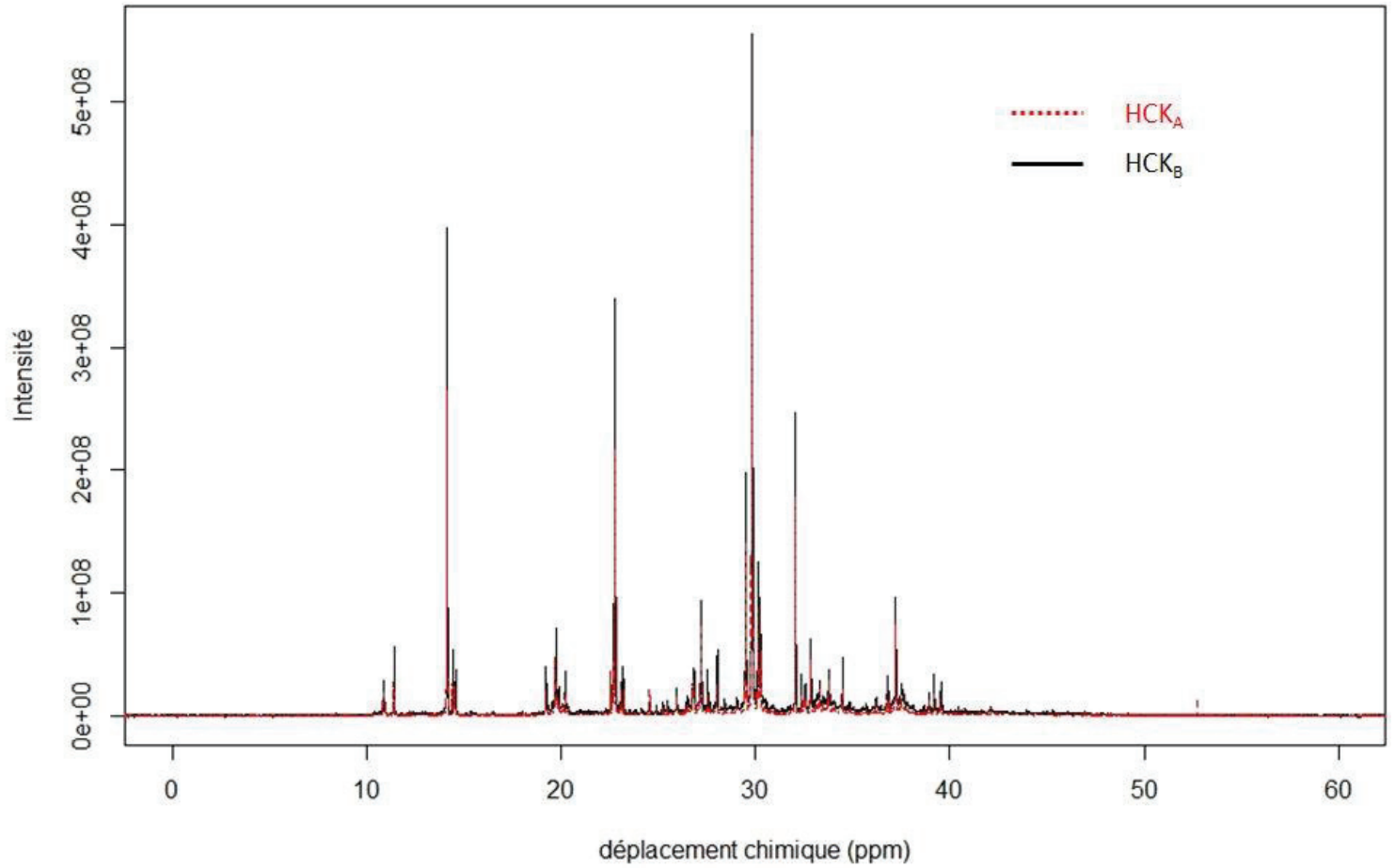


Figure 71 : Superposition des spectres ^{13}C RMN des échantillons d'huile produit avec HCK_A et HCK_B

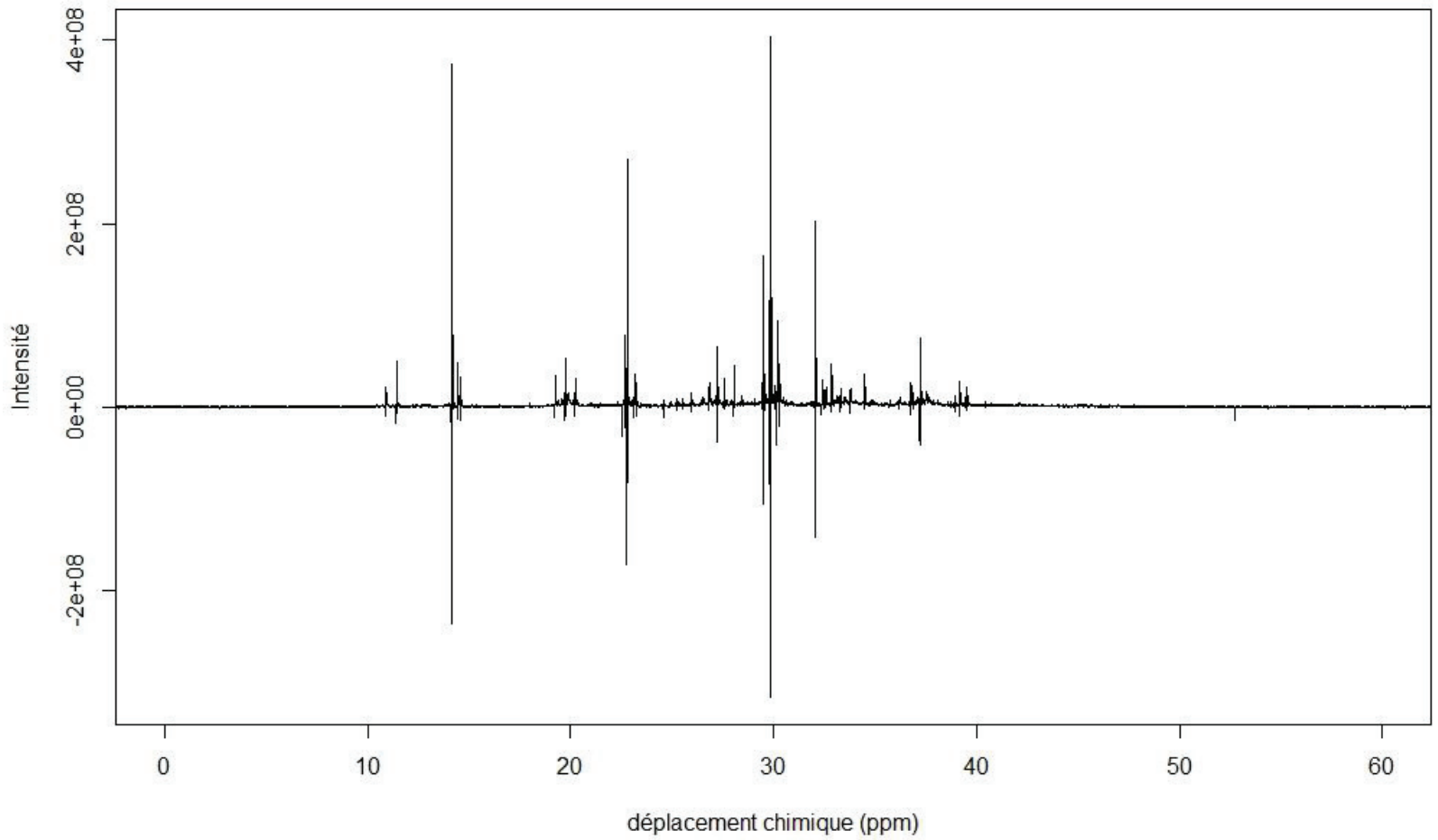


Figure 72 : Différence entre les deux spectres ^{13}C RMN des échantillons d'huile produit avec HCK_A et HCK_B

7.3.1.2 Exploitation de l'intégration des pics caractéristiques

Conformément à la méthodologie de comparaison des spectres ^{13}C RMN proposée au paragraphe 3.2.2.3, les pics significatifs des spectres des 4 échantillons à comparer ont été répertoriés, puis intégrés. Les valeurs moyennes et de fidélités associées au calcul des aires des pics caractéristiques pour les 4 échantillons d'huile étudiés sont données dans le Tableau 32. Le catalyseur utilisé pour la production de chaque échantillon est également spécifié. Les pics répertoriés dans ce tableau sont issus d'études antérieures [73,170,172] et sont également les plus significatifs de nos échantillons. Comme préconisé par Verdier *et al.* [73], l'aire du pic situé autour de 14,2 ppm a été prise comme référence et fixée à 1. Les valeurs données dans le Tableau 32 sont donc les aires des pics répertoriés normalisées par l'aire du pic de référence.

Pour exploiter ces résultats nous avons procédé comme suit :

1. Pour chaque pic, comparer les aires normalisées des échantillons produits respectivement avec HCK_A (5) et avec HCK_B (8).
2. Si une différence significative est notée, vérifier que les valeurs correspondantes pour les échantillons produits à partir de HCK_{B+A} (13) et HCK_{A+B} (17) sont intermédiaires et proches l'une de l'autre.

Cette procédure permet de vérifier d'une part qu'il y a ou non des différences entre les spectres, et d'autre part que ces différences suivent les mêmes tendances que le VI. Les colonnes correspondantes aux pics pour lesquels ces deux points sont vérifiés sont en gras (Tableau 32). Il s'agit :

- du pic à 11,5 ppm, caractéristique d'un méthyle (CH_3) à la fin d'un branchement éthyle (CH_2CH_3)
- du pic à 29,9 ppm, caractéristique d'un méthylène (CH_2) en milieu d'une chaîne droite.

La structure reliée au pic à 11,5 ppm est plus abondante dans l'échantillon 8 que dans l'échantillon 5 (respectivement $0{,}15 \pm 0{,}003$ et $0{,}11 \pm 0{,}006$). Cela signifie que **le catalyseur HCK_B à tendance à produire plus de branchement éthyle que le catalyseur HCK_A.** Le constat inverse peut être fait concernant le pic à 29,9 ppm. En effet, l'aire de ce pic est clairement plus importante pour l'échantillon 5 ($2{,}46 \pm 0{,}01$) que pour l'échantillon 8 ($1{,}87 \pm 0{,}02$). **L'abondance de chaîne linéaire est donc plus importante lorsque le catalyseur HCK_A (pouvoir craquant plus fort) est utilisé qu'avec le catalyseur HCK_B (pouvoir isomérisant). Ceci est logique car les étapes de craquage sont ordonnées comme suit :**

1. isomérisation
2. craquage des chaînes les plus ramifiées.

L'interprétation est donc la suivante : 1) le catalyseur HCK_A a un pouvoir craquant plus fort, on attend donc une abondance de chaînes linéaires (résultant d'un craquage de chaînes isomérisées). Ce qui est confirmé par la RMN. Les chaînes linéaires ayant un fort impact sur le VI, ce dernier est donc plus élevé ; 2) le catalyseur HCK_B a un pouvoir isomérisant plus fort. On attend donc une abondance de chaînes ramifiées (résultant d'une forte isomérisation). Ceci est confirmé par la RMN (pic à 11,5 ppm). Les chaînes isomérisées ayant un fort impact sur le VI, celui-ci est plus faible.

Bien qu'elles ne puissent pas être rigoureusement reliées à la variation du VI selon la procédure décrite ci-dessus, d'autres différences significatives entre HCK_A et HCK_B peuvent être notées. C'est le cas de l'aire des pics à 30,3 ppm, 32,0 ppm et 34,5 ppm. Le pic à 34,5 ppm, caractéristique d'un CH_2 lié à un méthine (CH) sur lequel est branché un groupe propyle, butyle, pentyle ou hexyle (Annexe E) est plus important dans l'échantillon produit avec le catalyseur HCK_B que dans celui qui a été obtenu avec le catalyseur HCK_A. Pour les deux autres (30,3 et 32,0 ppm) caractéristiques respectifs d'un CH_2 en position γ par rapport à un CH sur une chaîne droite lié à un branchement CH_3 et d'un CH_2 en position γ par rapport à un CH_3 au bout d'une chaîne droite (Annexe E), l'effet contraire est observé.

Globalement, cette comparaison de spectres ^{13}C RMN montre que HCK_B tend à favoriser des structures de branchement dans les molécules présentes dans les huiles, tandis que HCK_A produit majoritairement des molécules avec de longues chaînes alkyles et que ces différences ont un impact sur le VI.

7.3.1.3 Discussion par rapport à la composition des échantillons d'huile

L'étude du VI à partir des données GC×GC présentée au chapitre précédent a montré que les isoparaffines C_{28} à C_{46} ont une influence prépondérante sur le VI de la coupe huile. Les distributions en nombre de carbone des isoparaffines C_{28} à C_{46} ont donc été représentées sur la Figure 73 pour l'échantillon produit avec HCK_A et celui produit avec HCK_B. On observe une plus forte abondance de ce type de composé dans l'échantillon produit avec HCK_A. Cela est sans doute dû au pouvoir craquant de ce catalyseur. Le pouvoir isomérisant du catalyseur HCK_B ne produit pas plus d'isoparaffines mais certainement plus de branchements sur les molécules comme le montre la comparaison des spectres RMN (Figure 72).

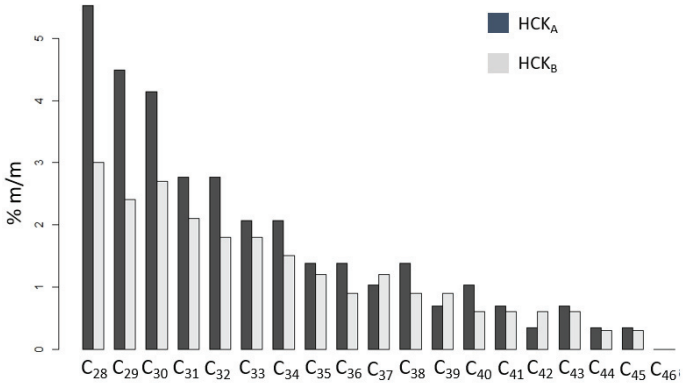


Figure 73 : Comparaison des distributions en nombre de carbone des isoparaffines pour les échantillons produits avec HCK$_A$ (5) et HCK$_B$ (8)

Tableau 32 : Résultats des intégrations des pics caractéristiques répertoriés sur les spectres RMN des échantillons d'huile

Echantillon (type de Catalyseur)	VI	Statistiques	Pic répertorié (ppm)										
			14,2[a]	**11,5**	19,9	22,8	26,8	27,5	**29,9**	30,3	32,0	34,5	37,0
Echantillon 5	118	Aire	1	**0,11**	0,41	0,62	0,31	0,08	**2,46**	0,77	0,56	0,10	0,33
(HCK_A)		IC		**0,006**	0,008	0,02	0,004	0,01	**0,01**	0,02	0,01	0,002	0,01
Echantillon 8	102	Aire	1	**0,15**	0,43	0,68	0,28	0,11	**1,87**	0,66	0,53	0,13	0,31
(HCK_B)		IC		**0,003**	0,01	0,07	0,03	0,01	**0,02**	0,02	0,01	0,006	0,02
Echantillon 13	112	Aire	1	**0,13**	0,41	0,66	0,27	0,09	**2,22**	0,67	0,55	0,11	0,34
(HCK_{B+A})		IC		**0,004**	0,01	0,01	0,003	0,007	**0,03**	0,02	0,01	0,01	0,004
Echantillon 17	108	Aire	1	**0,13**	0,43	0,65	0,28	0,11	**2,20**	0,69	0,58	0,14	0,30
(HCK_{A+B})		IC		**0,005**	0,01	0,01	0,01	0,02	**0,05**	0,04	0,01	0,04	0,01

[a]Pic de référence (Verdier *et al.* [73])

7.3.2 Comparaison des spectres d'échantillons de gazole

Une comparaison identique a été faite sur les 4 échantillons de gazole. Les valeurs moyennes et les fidélités des aires des pics caractéristiques répertoriés sont précisées dans le Tableau 33. Deux pics semblent clairement corrélés au PT :

- Le pic à 29,9 ppm, qui apparaissait déjà dans le cas du VI (méthylène au milieu une chaîne droite)
- Le pic à 34,5 ppm (caractéristique d'un méthylène en position α par rapport à un branchement propyle, butyle, pentyle ou hexyle).

Comme dans la coupe huile, l'abondance de méthylène en milieu d'une chaîne droite est plus forte dans la coupe gazole lorsque HCK_A est utilisé que lorsqu'il s'agit de HCK_B. De même, la structure de branchement liée au pic à 34,5 ppm est plus abondante cette fois avec le catalyseur HCK_B qu'avec le catalyseur HCK_A. Les effets des catalyseurs semblent donc aller dans le même sens quelle que soit la coupe observée. Enfin, les tendances montrent que plus l'aire du pic à 29,9 ppm est élevée, plus le PT augmente et *a contrario* que le PT diminue lorsque l'aire du pic à 34,5 ppm augmente.

D'autres différences significatives qui ne peuvent pas être directement reliées au PT ont été observées. Ces différences concernent les pics à 11,5 ppm, 19,9 ppm, 27,5 ppm et 37,0 ppm. Ces pics ont en commun d'être caractéristique d'une structure de branchement (voir Annexe E) et d'avoir une abondance plus forte dans l'échantillon de gazole produit avec HCK_B que dans celui qui a été obtenu avec HCK_A. Cela permet d'étendre la tendance globale relevée dans le cas de la coupe huile à celui de la coupe gazole.

7.3.2.1 Discussion par rapport à la composition des échantillons de gazole

Nous avons représenté ci-dessous (Figure 74) les distributions en n-paraffines et isoparaffines fonction du nombre de carbone pour les échantillons de gazole 1 et 6. Dans le cas des isoparaffines les distributions sont clairement équivalentes (Figure 74b). Pour les n-paraffines on observe une plus forte abondance des composés les plus lourds pour l'échantillon produit avec HCK_A (1) (Figure 74a). Pour ces échantillons, cette différence de teneur en n-paraffines lourdes semble suffire à expliquer la différence de PT. Par contre l'équivalence entre les teneurs en isoparaffines confirme que la différence entre les deux catalyseurs se situe au niveau de la structure pour ces composés.

Que ce soit pour la coupe huile ou pour la coupe gazole, les résultats de la comparaison des signaux ^{13}C RMN renforce l'intérêt de cette technique. Le lien entre le taux de branchement et la décroissance des propriétés d'intérêt est clairement démontré et renforce les précédents travaux de la littérature. De plus, les résultats de l'analyse RMN confirment également la différence d'isomérisation entre le catalyseur HCK_A et le catalyseur HCK_B, jusqu'à présent supposé mais jamais démontré au sein d'IFPEN.

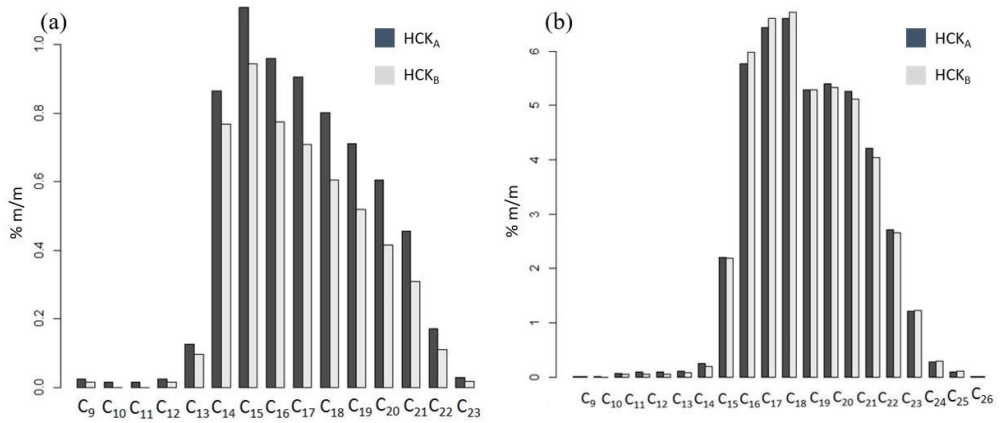


Figure 74 : Echantillons de gazole produits avec HCK_A (1) et HCK_B (6) ; a) Comparaison des distributions des n-paraffines en fonction du nombre de carbone ; b) Comparaison des distributions des isoparaffines en fonction du nombre de carbone

Tableau 33 : Résultats des intégrations des pics caractéristiques répertoriés sur les spectres RMN des échantillons de gazole

Catalyseur utilisé	PT (°C)	Statistiques sur les aires	Pic répertorié (ppm)										
			14,2[a]	11,5	19,9	22,8	26,8	27,5	**29,9**	30,3	32,0	**34,5**	37,0
Echantillon 1 (HCK_A)	-19	Aire	1	0,20	0,23	0,69	0,19	0,13	**1,41**	0,39	0,59	**0,15**	0,22
		IC		0,01	0,01	0,01	0,006	0,01	**0,01**	0,02	0,02	**0,01**	0,01
Echantillon 6 (HCK_B)	-30	Aire	1	0,23	0,26	0,67	0,20	0,16	**1,23**	0,43	0,56	**0,19**	0,24
		IC		0,01	0,01	0,02	0,01	0,01	**0,05**	0,02	0,02	**0,01**	0,01
Echantillon 11 (HCK_{B+A})	-24	Aire	1	0,22	0,26	0,69	0,20	0,17	**1,37**	0,44	0,59	**0,17**	0,24
		IC		0,01	0,002	0,02	0,004	0,004	**0,01**	0,005	0,01	**0,003**	0,004
Echantillon 16 (HCK_{A+B})	-25	Aire	1	0,22	0,27	0,67	0,20	0,15	**1,32**	0,44	0,58	**0,18**	0,24
		IC		0,003	0,004	0,004	0,003	0,004	**0,02**	0,01	0,01	**0,004**	0,005

[a]Pic de référence (Verdier *et al.* [73])

7.4 Conclusion

L'application de la *sparse* PLS aux données issues de l'analyse RMN du^{13}C de la coupe huile et de la coupe gazole a permis d'identifier un certain nombre de marqueurs moléculaires qui sont fortement reliés aux propriétés d'intérêt. Dans le cas des huiles, la méthodologie a montré des tendances déjà relevées dans la littérature. La qualité de la modélisation du VI est d'autant plus satisfaisante que l'étude analogue à partir des données GC×GC n'avait pas fourni de résultats probants. Dans le cas des gazoles, l'analyse multivariée de signaux ^{13}C RMN n'avait que très peu été utilisée auparavant pour la compréhension des propriétés à froid. En effet, la caractérisation des gazoles par des techniques chromatographiques telles que la GC×GC fournissait des résultats suffisants. Or, la tendance du PT à croître avec la présence de méthylène concorde avec l'effet des n-paraffines qui sont les molécules qui comportent le plus gros ratio de ce type de carbone. De la même manière, les structures de branchements qui tendent à diminuer le PT coïncident avec l'impact des isoparaffines. Cette synthèse souligne notamment la complémentarité des deux techniques de caractérisation.

Outre la capacité de la *sparse* PLS à faire ressortir les zones spectrales d'intérêt pour une propriété donnée, la qualité des modèles obtenus ouvre la perspective de prédire le PT de la coupe gazole à partir de signaux ^{13}C RMN d'effluents totaux.

Enfin, l'étude de comparaison de spectre RMN du ^{13}C a mis en évidence des différences significatives entre les spectres des échantillons produits avec des catalyseurs différents aussi bien dans le cas des huiles que dans le cas des gazoles. Ces différences traduisent une nette tendance du catalyseur HCK_B à favoriser les structures de branchement par rapport au catalyseur HCK_A, dont certaines influent clairement sur le VI et le PT.

Chapitre 8. Prédiction des propriétés produits

Les travaux de compréhension moléculaire du PT de la coupe gazole et du VI de la coupe huile ont mis en évidence l'influence de certaines espèces d'hydrocarbures et de leur structure moléculaire sur ces propriétés. L'application des méthodes d'analyse multivariée telles que la PLS et *sparse* PLS fournit des modèles d'une précision satisfaisante. Ils ont également montré la faisabilité de la prédiction de ces propriétés à partir d'analyse de l'effluent total dans lequel les coupes gazole et huile sont contenues avant distillation. Rappelons que dans le simulateur HCK développé par IFPEN, seules les propriétés de base contenues dans les bilans expérimentaux peuvent être potentiellement utilisées comme descripteurs pour les modèles. Bien que les données chromatographiques ou spectroscopiques ne puissent pas être utilisées dans ces simulateurs de procédé, l'identification des marqueurs moléculaires qui influent sur les propriétés d'intérêt peut contribuer à améliorer la pertinence dans le choix des descripteurs. Par exemple, l'abondance des longues chaînes alkyles, qui comme on l'a vu influe sur le PT et sur le VI, est directement reliée aux points de coupe (obtenus par distillation simulée). De la même manière, la présence de carbones isoparaffiniques dépendra du taux de conversion appliquée. Au cours de notre étude, plusieurs modèles ont été développés dans le but de proposer des perspectives d'amélioration des performances du simulateur HCK. L'objet de ce chapitre est de présenter et discuter les performances des modèles proposés pour la prédiction du PT de la coupe gazole et du VI de la coupe huile à partir de propriétés de base des différentes coupes mises en jeu durant le procédé d'HCK. Nous présenterons en premier lieu l'étude qui a été menée autour du PT de la coupe gazole. Ensuite, nous reviendrons sur les travaux analogues qui ont été réalisés pour la prédiction du VI de la coupe huile à partir de propriétés globales du liqtot.

8.1 Prédiction du PT à partir de propriétés de base

Dans ce paragraphe, les différents travaux qui ont été effectués pour prédire le PT de la coupe gazole sont détaillés. Les différents modèles qui ont été développés sont comparés et les résultats sont commentés et discutés.

Les descripteurs du modèle ont été choisis parmi toutes les propriétés de base (environ 80) contenues dans les 57 bilans retenus pour cette étude (paragraphe 5.3.1). Cette sélection a été effectuée sur deux critères :

1. un premier avis expert basé sur la connaissance du procédé d'hydrocraquage couplé aux résultats de l'étude de compréhension moléculaire du PT réalisée dans les chapitres 7 et 8,
2. le calcul des indices de sensibilité des propriétés retenues dans l'étape 1 (paragraphe 4.2.2).

Trois propriétés de base ont ainsi été sélectionnées pour le développement des modèles de prédiction du point de trouble de la coupe gazole :

1. le taux de conversion X_{370} (paragraphe 5.1.1.3),
2. la température T_{95} obtenue par DS (paragraphe 3.1.2),
3. la teneur en azote contenu dans la charge DSV d'origine.

La base de données a été divisée en deux : une base d'apprentissage constituée de 40 bilans et une base de test contenant 17 bilans. La projection des données dans le plan (X_{370}, T_{95}) est illustrée sur la Figure 75. Pour faciliter la discussion, les échantillons ont été numérotés de 1 à 40 pour la base d'apprentissage, et de 41 à 57 pour la base de test. On peut noter que le PT diminue globalement lorsque X_{370} augmente et qu'à l'inverse, il augmente avec la T_{95}. Ces deux observations sont cohérentes avec les résultats de l'étude de compréhension moléculaire. En effet, l'augmentation de X_{370} se traduit par une plus forte isomérisation des molécules et donc une diminution du rapport de la teneur en n-paraffines sur la teneur en isoparaffines qui est un facteur de diminution du PT (paragraphe 6.1.2). De même, une augmentation de la T_{95} entraîne la présence de n-paraffines plus lourdes et favorise donc une cristallisation rapide. La teneur en azote de la charge d'origine n'est pas représentée ici mais traduit l'importance de la qualité la charge sur les réactions qui ont lieu durant le procédé. Elle confirme par ailleurs l'influence de celle-ci sur la composition moléculaire de la coupe gazole tel qu'on a pu l'observé dans les chapitres précédents.

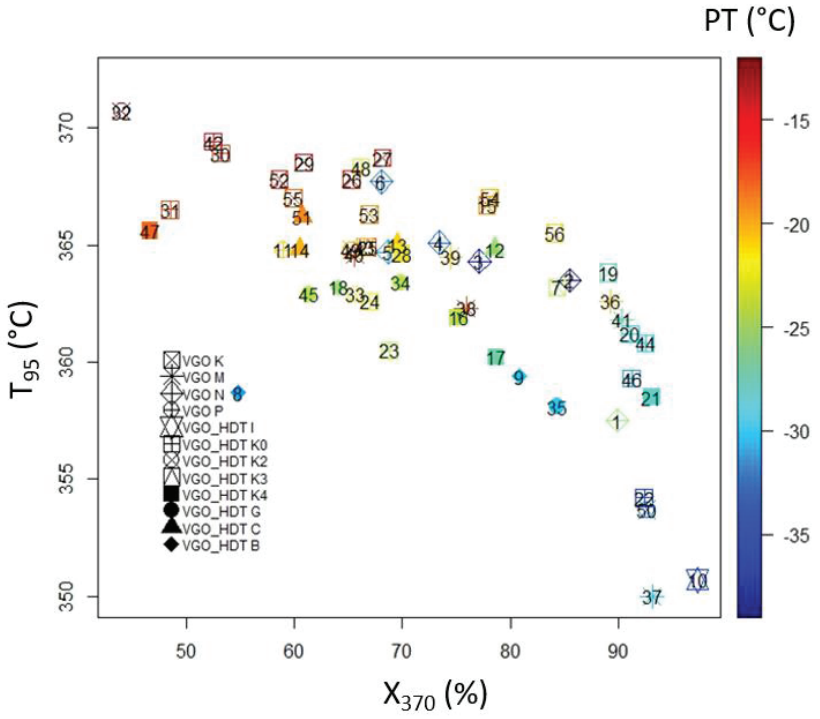


Figure 75 : Projection de la base de données de gazole sur le plan (X_{370}, T_{95})

Un modèle RLM et un modèle de krigeage ont été développés en utilisant essentiellement la base d'apprentissage. Les résultats sont présentés ci-dessous.

8.1.1 Comparaison des performances des modèles

Les performances des modèles RLM et de krigeage ont été évaluées sur les données d'apprentissage par méthode *leave-one-out* (paragraphe 4.6), puis sur la base de test. A titre comparatif, le modèle actuel du simulateur IFPEN pour la prédiction du PT a été recalibré à partir de nos données d'apprentissage et ses performances ont été évaluées suivant la même procédure que nos modèles.

Les graphes de parité obtenus pour le modèle RLM et le modèle de krigeage sur les données d'apprentissage sont illustrés sur la Figure 76. La droite de parité est représentée en rouge (trait plein), les droites en pointillés verts et bleus délimitent respectivement les intervalles de confiance à 95% (précision à ±IC) et à 68% (précision à ±2IC) associés à la mesure de référence ($IC = 2{,}8°C$ dans le cas du PT). Enfin, les intervalles de confiance de prédiction sont illustrés pour chaque point par des barres d'erreurs. On note que dans certains cas la valeur prédite peut être significativement différente selon le modèle considéré. Par exemple l'échantillon 32 est bien prédit dans le cas du modèle RLM puisqu'il est clairement localisé à l'intérieur de l'IC de la mesure de référence (Figure 76a) tandis que son estimation par krigeage est assez éloignée de la mesure (Figure 76b). L'observation inverse peut être faite dans le cas de l'échantillon 2. Globalement, la précision semble meilleure dans le cas du krigeage que dans celui du modèle RLM. Ce constat est appuyé par les statistiques des modèles qui sont précisées dans le Tableau 34. En effet, une RMSECV de 4,0°C a été obtenue dans le cas du modèle RLM contre 2,9°C pour le modèle de krigeage (soit une différence de 1,0°C). De plus, le pourcentage de points prédits dans l'IC de la mesure de référence est plus élevé dans le cas du krigeage (70%) que dans le cas de la RLM (60%). Toujours selon la méthode *leave-one-out*, les modèles que nous proposons ont tous deux des statistiques bien meilleures que celles du simulateur HCK actuel (Tableau 34). Cette fois la RMSECV obtenue est de 5,6 et 42% des points sont prédits dans l'IC de la mesure.

Dans le cas du modèle RLM, l'amplitude des barres d'incertitudes associées aux valeurs prédites est naturellement quasi-homogène. Dans le cas du krigeage ces barres d'erreurs fournissent des informations intéressantes. En effet, on note que les échantillons 8, 10, 22 et 37, dont les incertitudes de prédiction sont les plus fortes (barres d'erreur les plus longues, Figure 76b) sont relativement éloignés (individuellement) du reste de la base d'apprentissage (Figure 75). Ce résultat renforce l'intérêt des incertitudes stochastiques qui permettent de discuter de la valeur prédite en un point suivant sa position par rapport aux données d'apprentissage.

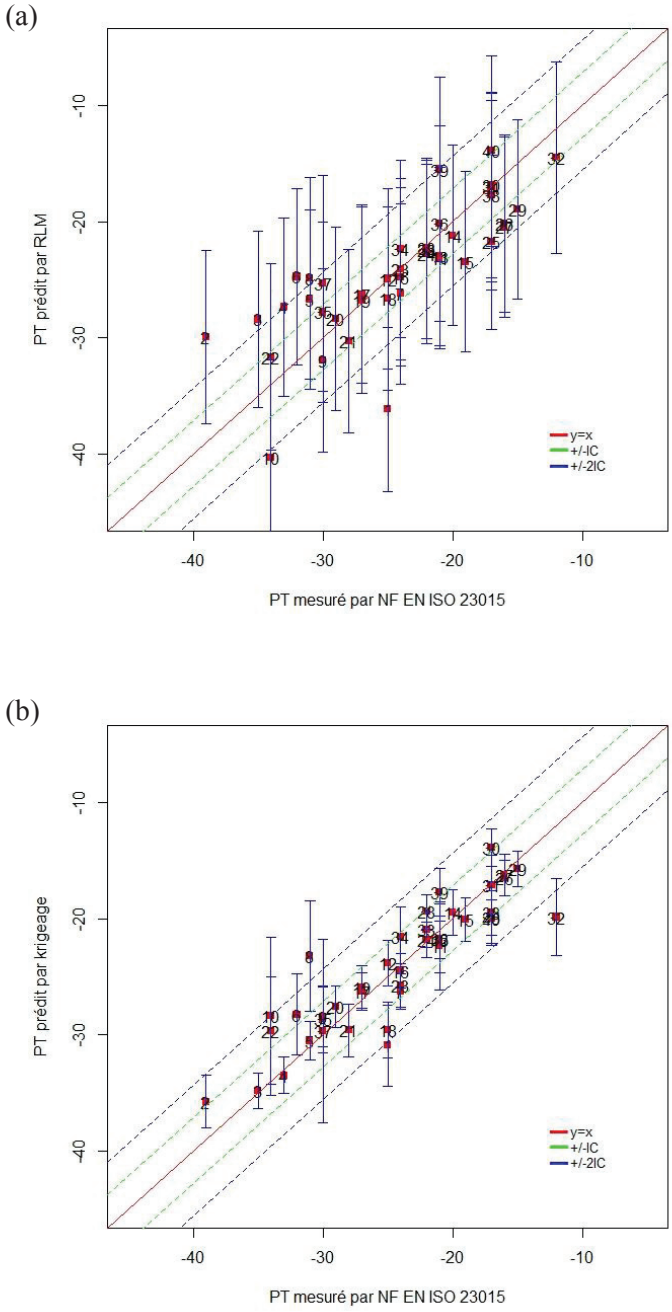


Figure 76 : Graphes de parité des modèles de prédiction du PT par *leave-one-out* sur la base d'apprentissage. a) modèle RLM ; b) modèle de krigeage

Les modèles ont également été évalués sur la base de test. Les graphes de parité correspondant sont représentés sur la Figure 77. Comme pour les données d'apprentissage, le modèle de krigeage semble globalement plus précis que le modèle RLM (Figure 77b et Figure 77a respectivement). Cette tendance est là aussi confirmée par les statistiques de performances RMSEP de 2,0°C a été obtenue avec le modèle de krigeage contre 2,5°C dans le cas du modèle RLM. De plus, 88% des échantillons de la base de test sont prédits avec une erreur inférieure à l'IC de la mesure pour le modèle krigé contre 70%

pour le modèle RLM. Les performances du modèle IFPEN actuel sont encore globalement moins bonnes que ceux des modèles proposés puisque seulement 56% des points sont prédits dans l'IC de référence et une différence de 1°C entre la RMSEP obtenue et celle du modèle krigé est observée. On peut noter que dans certains cas la qualité de la prédiction est meilleure pour le modèle RLM par rapport au modèle de krigeage (points 41 et 48, Figure 77).

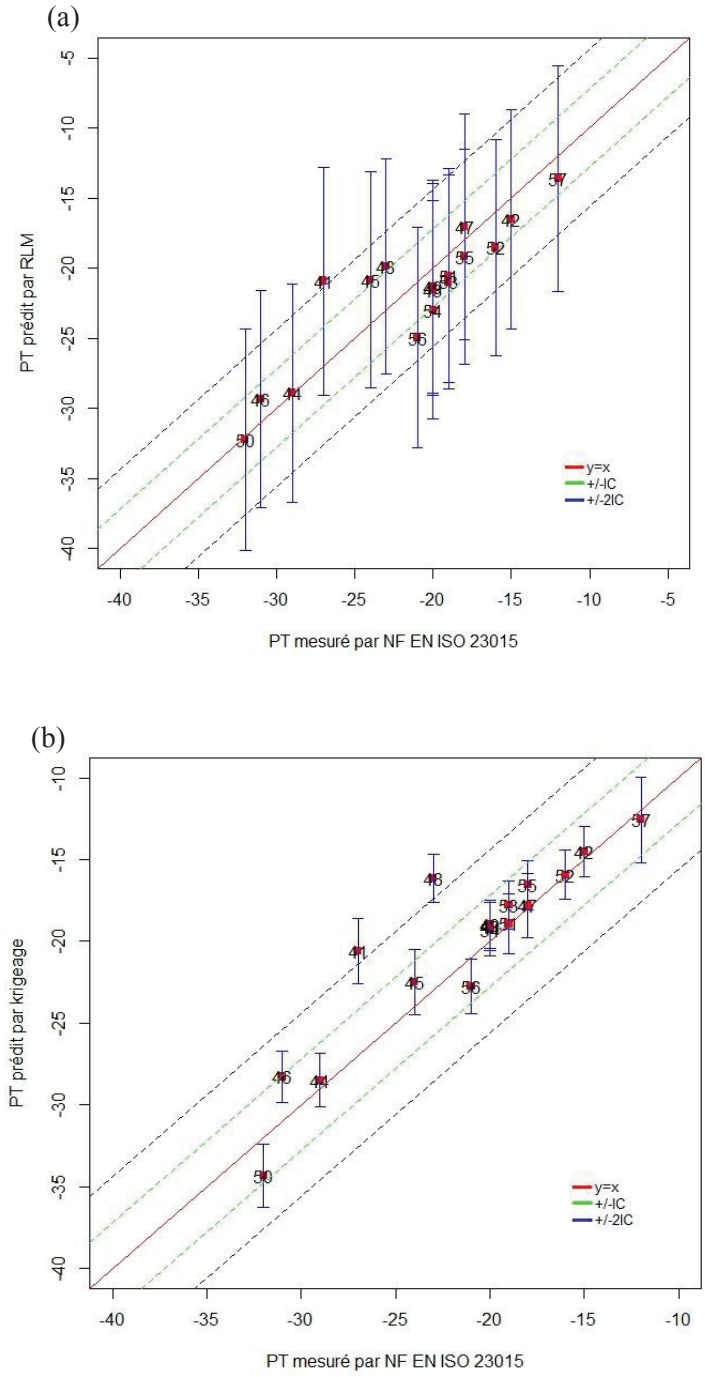


Figure 77 : Graphes de parité des modèles de prédiction du PT sur la base de test ; a) modèle RLM ; b) modèle de krigeage

Tableau 34 : Statistiques des modèles de prédiction du PT de la coupe gazole à partir des propriétés globales de la coupe

Base d'évaluation	Modèle	MAD (°C)	RMSE* (°C)	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
Apprentissage (*Leave-one-out*)	RLM	3,0	4,0	60	82,5
	Krigeage	2,1	2,9	70	90
	Simulateur IFPEN actuel	4,3	5,6	42,5	72,5
Test	RLM	2,0	2,5	70	94
	Krigeage	1,7	2,0	88	88
	Simulateur IFPEN actuel	2,6	3,0	56	100

IC → Intervalle de confiance de la méthode NF EN ISO 23015 ; * → RMSECV pour la base d'apprentissage et RMSEP pour la base de test.

Globalement, on note que le modèle de prédiction du PT de la coupe gazole par krigeage est plus performant que le modèle RLM avec les mêmes descripteurs. Ce résultat souligne la capacité de cette méthode à améliorer la prédiction dans lorsque la propriété à modéliser présente des caractéristiques non linéaires. De plus, la comparaison des performances des modèles proposés à ceux du modèle du simulateur IFPEN sur une même base de données met en avant la pertinence des descripteurs sélectionnés dans notre étude. Tout comme le PT de la coupe gazole, le VI de la coupe huile présente lui aussi des effets de non linéarité. Les résultats du modèle de krigeage présentés dans ce paragraphe renforcent la perspective de l'utilisation de cette méthode pour la prédiction du VI.

8.2 Prédiction du VI de la coupe huile

Des modèles de prédiction du VI de la coupe huile ont été développés selon une démarche analogue à celle qui a été utilisée pour la prédiction du PT. Dans ce paragraphe, nous présentons et commentons les performances de ces modèles. Deux cas de prédiction du VI ont été traités en parallèle : le premier concerne uniquement les huiles issues de tests HDT ; le second fait référence essentiellement aux huiles provenant de tests HCK (paragraphe 5.3.2). **Rappelons que dans ces deux cas, les données analytiques qui sont utilisées comme descripteurs ont été mesurées sur le liqtot.** Ce choix a été introduit au paragraphe 2.5.3.

Comme dans le cas du PT de la coupe gazole, le choix des descripteurs a été fait en deux étapes :

1. présélection par connaissance des procédés couplée aux conclusions des travaux de compréhension moléculaire du VI,
2. sélection des descripteurs finaux par maximisation du $R^2_{ajusté}$ (paragraphe 4.2.1).

A chaque fois, la base de données a été divisée aléatoirement en une base d'apprentissage et une base de test.

8.2.1 Prédiction du VI de la coupe huile issue de tests HDT

8.2.1.1 Visualisation des données par ACP et choix des descripteurs

Une ACP a été réalisée sur l'ensemble de la base de données en tenant compte uniquement des potentiels descripteurs présélectionnés dans le but de vérifier la pertinence de ces choix. La Figure 78a représente les scores des données caractéristiques des bilans HDT sur le premier plan (PC1, PC2) qui représente 92% de la variance expliquée. Le VI de la coupe huile correspondante est représenté par une échelle de couleur (rouge pour les haut VI et bleu pour les VI bas). Le cercle de corrélations des variables explicatives sur le plan (PC1, PC2) est représenté sur la Figure 78b. On observe la présence de clusters caractéristiques d'échantillons provenant d'une même charge DSV ou de charges ayant des propriétés très voisines. On note également une variation monotone du VI dans le plan (PC1, PC2). Cela implique d'une part la nécessité d'introduire dans les modèles des informations relatives à la charge d'origine et d'autre part que le VI est bien expliqué par PC1 et PC2. L'importance de la charge d'origine avait déjà été souligné au cours de l'étude de compréhension moléculaire du VI qui montrait clairement son influence prépondérante sur la composition des coupes pétrolières (Chapitre 6 et Chapitre 7).

Le cercle de corrélation donne les coefficients de corrélation entre les variables d'une part, puis entre les variables et les composantes principales d'autre part. Le coefficient de corrélation d'une variable à PC1 (respectivement PC2) est obtenu par projection orthogonale de la variable sur l'axe horizontal (respectivement vertical). De plus, deux variables dont les projections sont très proches sur le cercle de corrélation sont inter-corrélées. Si leurs projections sont diamétralement opposées alors elles sont anti-corrélées.

La Figure 78b montre que PC1 est très corrélée à la proportion en carbone paraffinique (C_p), à l'API Gravity et au facteur de caractérisation de Watson (K_w), et anti-corrélée à la densité (d_{15}^4), l'indice de réfraction (IR) et la proportion en carbones aromatiques (C_a). Ainsi PC1 renvoie à des informations sur la densité des liqtot issus de bilans HDT. PC2 est quant à elle corrélée aux propriétés relatives à la volatilité des échantillons de la base de données ($MeABP$, T_M, T_{MP}) ainsi qu'à leur masse molaire (M). On note de plus que K_w qui peut être interprété comme un ratio point d'ébullition/densité est la variable la plus corrélée au VI. Les observations ci-dessous peuvent être traduites comme suit : dans le cas de l'HDT des DSV, les effluents totaux à partir desquels les coupes huile de plus haut VI sont obtenues sont ceux qui ont à la fois un point d'ébullition élevé et une densité faible. Ces caractéristiques sont propres aux composés paraffiniques et contraire à celles des composés aromatiques et traduisent donc une forte abondance des premiers et une faible abondance des seconds. Ces conclusions sont en adéquation avec les corrélations proposées par Sarpal *et al.* [109] et par Sharma *et al.* [117]. Cette explication peut cependant paraître simpliste car les composés normal paraffiniques et iso paraffiniques

qui constituent l'ensemble des paraffines peuvent avoir des propriétés rhéologiques très diverses selon leur structure. La tendance des composés aromatiques à être néfaste pour le VI a également été soulignée par Verdier *et al.* [73] et Sastry *et al.* [171]. Elle est de plus en accord avec les conclusions de l'étude compréhension moléculaire du VI à partir des données GC×GC (paragraphe 6.2.3.2).

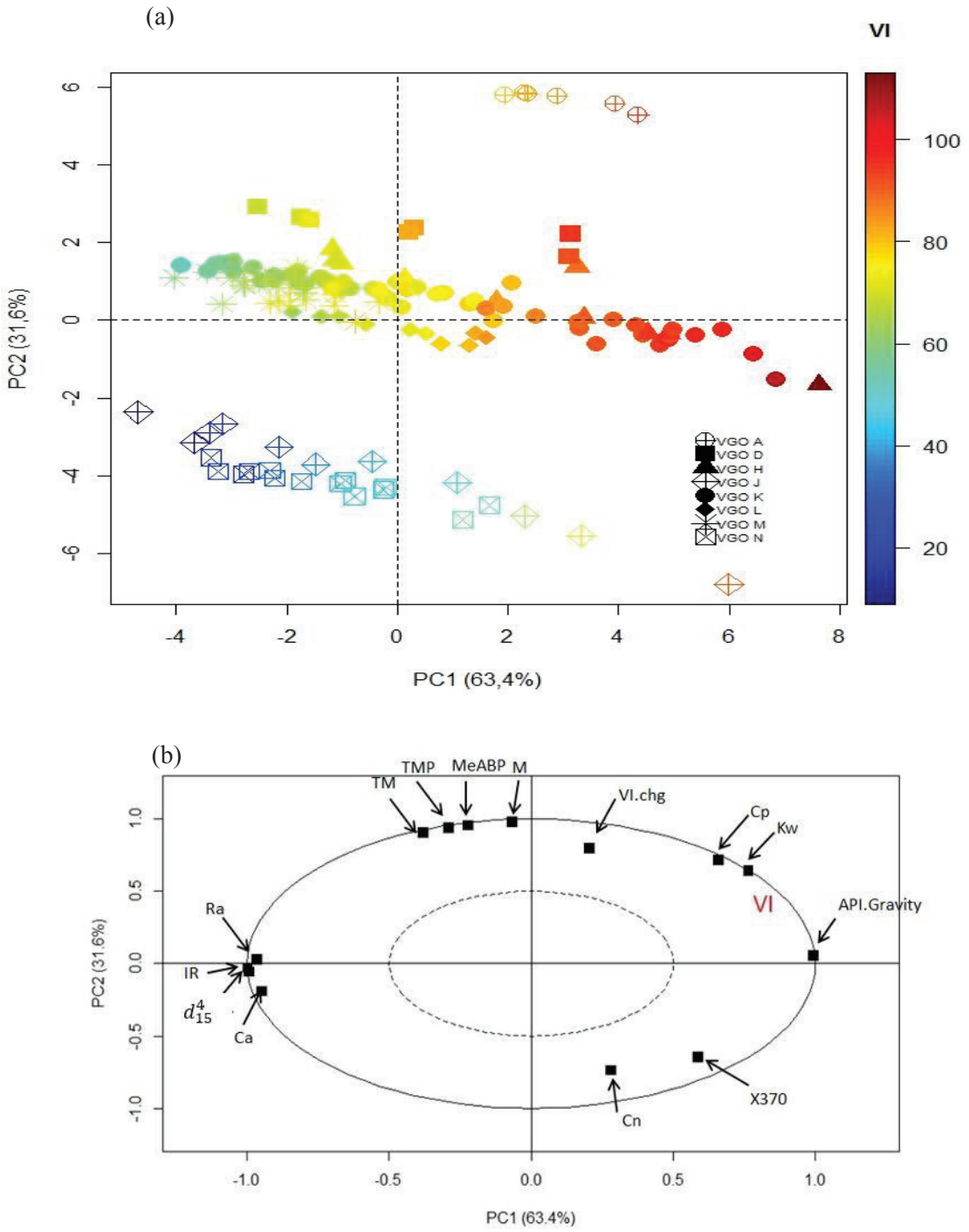


Figure 78 : a) Scores des bilans de liqtot HDT sur le premier plan factoriel ; b) Cercle de corrélation des variables

Les observations obtenues à partir de l'ACP permettent de relier la variabilité des données aux phénomènes physico-chimiques qui ont lieu durant l'étape d'HDT. Cela donne une vraie crédibilité au choix des variables mises en jeu. A la suite de cette ACP, une sélection de variables a été réalisée par maximisation du $R^2_{ajusté}$ (paragraphe 4.2.1). Six variables ont été retenues : d^4_{15}, MeABP, M, K_w,VI de la charge ($VI.chg$) (voir Tableau 9) et X_{370}.

8.2.1.2 Comparaison des modèles de prédiction du VI

Un modèle RLM et un modèle de krigeage ont donc été développés pour la prédiction du VI à partir de ces descripteurs. Les graphes de parité des modèles sur la base d'apprentissage (90 bilans) obtenus par méthode *leave-one-out* sont représentés sur la Figure 79. La légende utilisée est la même que celle qui a été utilisée pour les modèles de prédiction du PT (paragraphe 8.1.1), les droites en pointillés vert et bleu représentant cette fois les intervalles de confiance liés à la méthode de mesure du VI (NF EN ISO 2909 [19]). Etant donné le domaine de VI couvert, nous avons fixé l'IC de la mesure de référence à 2 [12].

Ces graphes de parité sont très proches l'un de l'autre (Figure 79a et Figure 79b). Les statistiques des modèles sur la base d'apprentissage sont précisées dans le Tableau 35. Des valeurs de RMSECV de 2,4 et 2,0 ont été obtenues respectivement pour le modèle RLM et pour le modèle krigé. Le pourcentage de points prédits avec une erreur inférieure à l'IC de la méthode de mesure est de 63% pour le modèle RLM contre 77% pour le modèle de krigeage. Les performances de ce dernier sont donc légèrement meilleures que celles du modèle RLM sur la base d'apprentissage.

Le modèle IFPEN actuel a été recalibré à partir de nos données d'apprentissage. Il a été évalué suivant le même mode que nos modèles. Les statistiques montrent une très nette différence en faveur de nos modèles. En effet, la RMSECV obtenue pour le modèle IFPEN actuel est de 5,0 et seulement 37% des points sont prédits à l'intérieur de l'IC de la mesure.

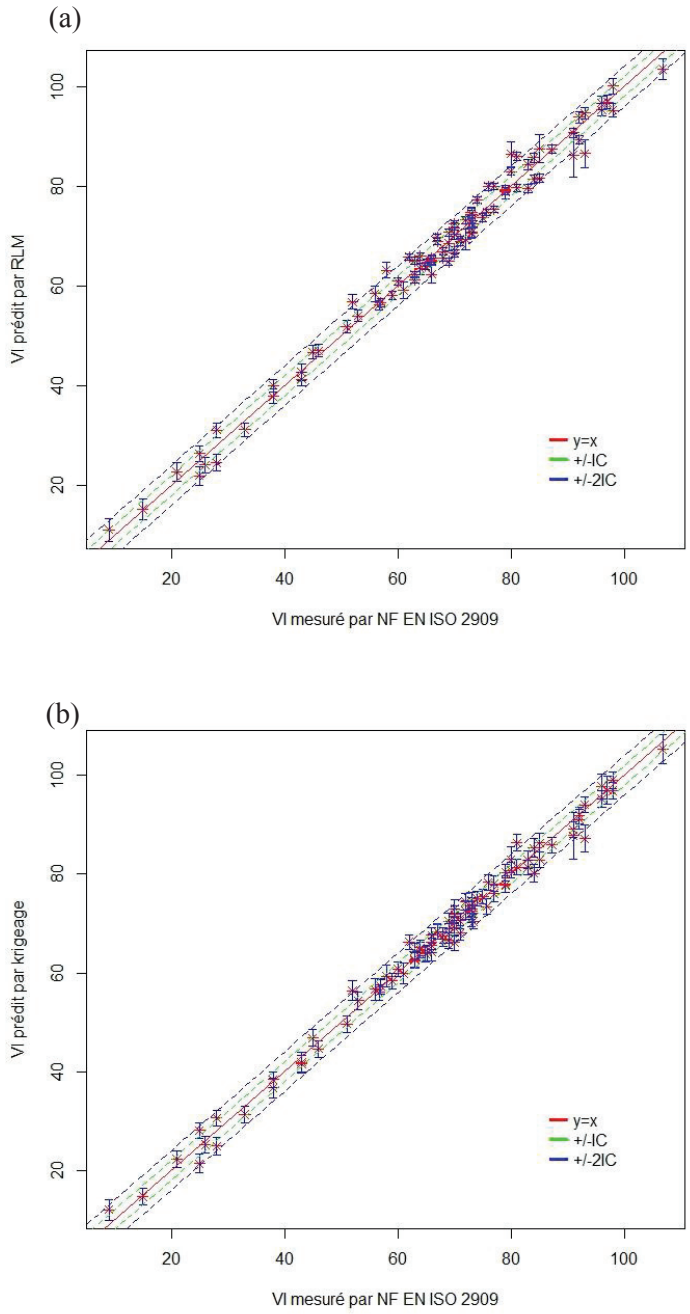


Figure 79 : Graphes de parité des modèles de prédiction du VI de la coupe huile issue de l'HDT des DSV par *leave-one-out* sur la base d'apprentissage. a) modèle RLM ; b) modèle de krigeage

Une comparaison similaire a été faite sur la base de test (45 bilans). Les graphes de parité des modèles correspondants sont représentés sur la Figure 80. Comme pour la base d'apprentissage, les modèles semblent très proches. Comparé au modèle RLM, le krigeage fournit là aussi des performances légèrement meilleures : le pourcentage $\tau_{\pm IC}$ est en faveur du krigeage (68% contre 63%, Tableau 35), de même que les valeurs de RMSE (2,1 contre 2,4).

Les résultats précédents montrent que le modèle RLM et le modèle de krigeage sont relativement proches avec des prédictions légèrement meilleures pour le modèle krigé. Globalement, les deux modèles ont de bonnes performances. La bonne qualité du modèle RLM dans ce cas s'explique par l'absence d'isomérisation et de craquage durant l'HDT. La structure des molécules n'est pas modifiée (excepté pour les aromatiques). Il y a donc une vraie proximité entre la composition du liqtot et celle des coupes obtenues après distillation.

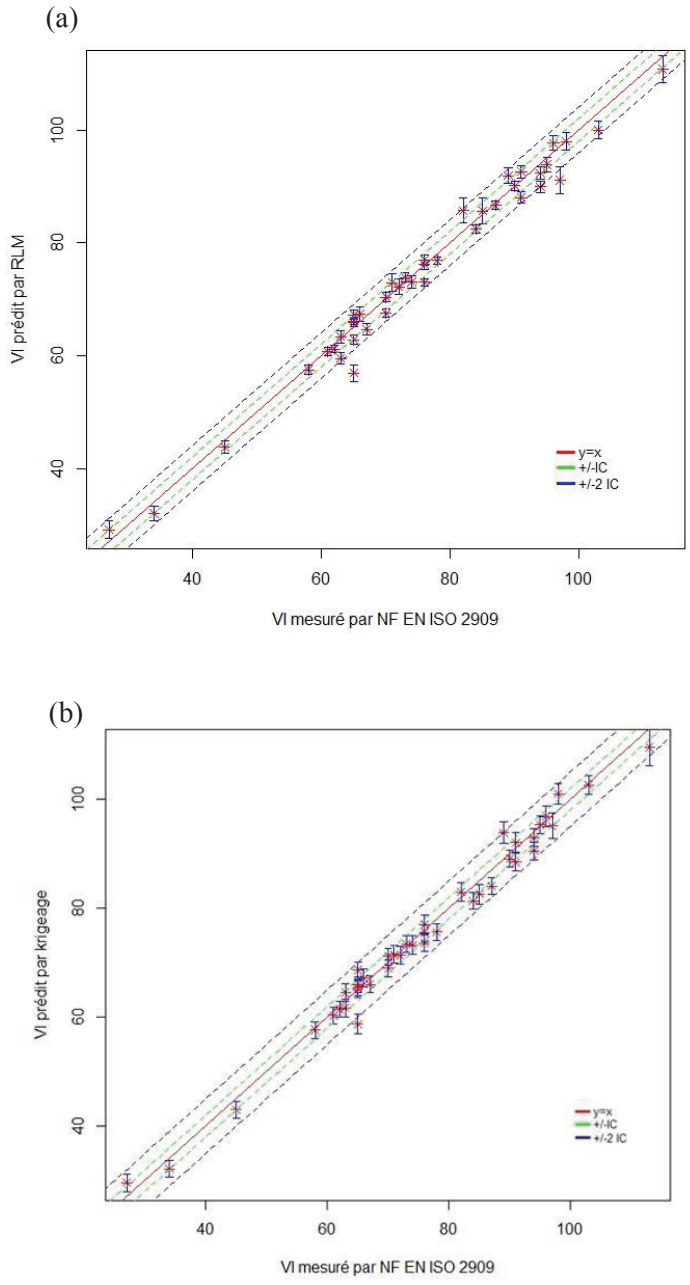


Figure 80 : Graphes de parité des modèles de prédiction du VI de la coupe huile issue de l'HDT des DSV sur la base de test. a) modèle RLM ; b) modèle de krigeage

De même que sur la base d'apprentissage, nous avons comparé les performances de nos modèles à celles du simulateur IFPEN actuel. Là encore, la différence est nette puisque seulement 10% des points sont prédits à l'intérieur de l'IC de la mesure. La RMSEP est de 5,3, soit 3 points de plus que le modèle krigé**. Globalement, les modèles proposés sont clairement plus performants que le modèle IFPEN actuel.**

Tableau 35 : Statistiques des modèles de prédiction du VI de la coupe huile issue de l'HDT des DSV à partir de propriétés globales du liqtot

Base d'évaluation	Modèle	MAD	RMSE	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
Apprentissage (*Leave-one-out*)	RLM	2,0	2,4	61	93
	Krigeage	1,9	2,0	77	98
	Simulateur IFPEN actuel	3,7	5,0	37	69
Test	RLM	1,8	2,4	63	92
	Krigeage	1,6	2,1	68	95
	Simulateur IFPEN actuel	4,6	5,3	10	60

IC →intervalle de confiance de la méthode NF EN ISO 2909 [169]; * → RMSECV pour la base d'apprentissage et RMSEP pour la base de test.

8.2.2 Prédiction du VI de la coupe huile issue de tests HCK

8.2.2.1 Visualisation des données par ACP et choix des descripteurs

Comme dans le cas des huiles issues de tests HDT, une ACP a été réalisée pour vérifier la pertinence de nos des descripteurs potentiels. Les scores des données de la base HCK et le cercle de corrélation des variables sur PC1 (80% de la variance totale expliquée) et PC2 (11%) sont représentés sur la Figure 81a et Figure 81b respectivement. Les observations sont très différentes de celles faites sur la base de données HDT. En effet, aucune tendance globale n'est observable concernant le VI de la coupe huile correspondante (Figure 81a) et on ne distingue la présence d'aucun cluster. Deux points apparaissent relativement éloignés des autres mais ils ne présentent pas d'anomalie particulière.

PC1 est maintenant corrélée à la d_{15}^4, l'IR, C_n ainsi qu'aux propriétés relatives au point d'ébullition (T_M, $MeABP$, *etc.*) (Figure 81b). PC1 est par ailleurs anti-corrélée à C_a, C_p, le VI de la charge DSV hydrotraitée d'origine et le taux de conversion (X_{370}). Bien que le VI de la coupe huile ne soit pas aussi bien expliqué dans ce cas que dans le précédent, on note qu'il est assez corrélé à PC1 et plus ou moins anti-corrélé à PC2. Ces observations peuvent être interprétées comme suit : la baisse simultanée des concentrations en aromatiques et en n-paraffines dans le liqtot (comparé à l'effluent total issu de l'étape HDT) dues à la conversion des premiers en naphtènes (réactions d'HDA) et à l'isomérisation des seconds lors de l'étape HCK (voir Annexe B) réduit significativement leur effet sur la densité et le point d'ébullition. En conséquence, ces deux propriétés sont plutôt liées aux composés naphténiques et isoparaffiniques présents en fortes proportions dans le liqtot en sortie HCK.

Deux variables sont particulièrement corrélées au VI : le VI de la charge DSV hydrotraité d'origine ($VI.chg$) et Cp. Dans le premier cas, cette observation confirme l'influence prépondérante de la charge qui a déjà été évoquée dans le cas des huiles issues de tests HDT. Dans le second cas, la variable Cp est très liée à la présence de paraffines et principalement d'isoparaffines qui sont nettement plus présentes dans la fraction huile. Son influence sur le VI est donc en accord les conclusions de l'étude du VI à partir des données GC×GC qui avait montré la tendance des isoparaffines à augmenter le VI. On note également que C_n est anti-corrélée au VI ce qui traduit une influence négative des carbones naphténiques qui avait déjà été soulignée par Sarpal *et al*. [109].

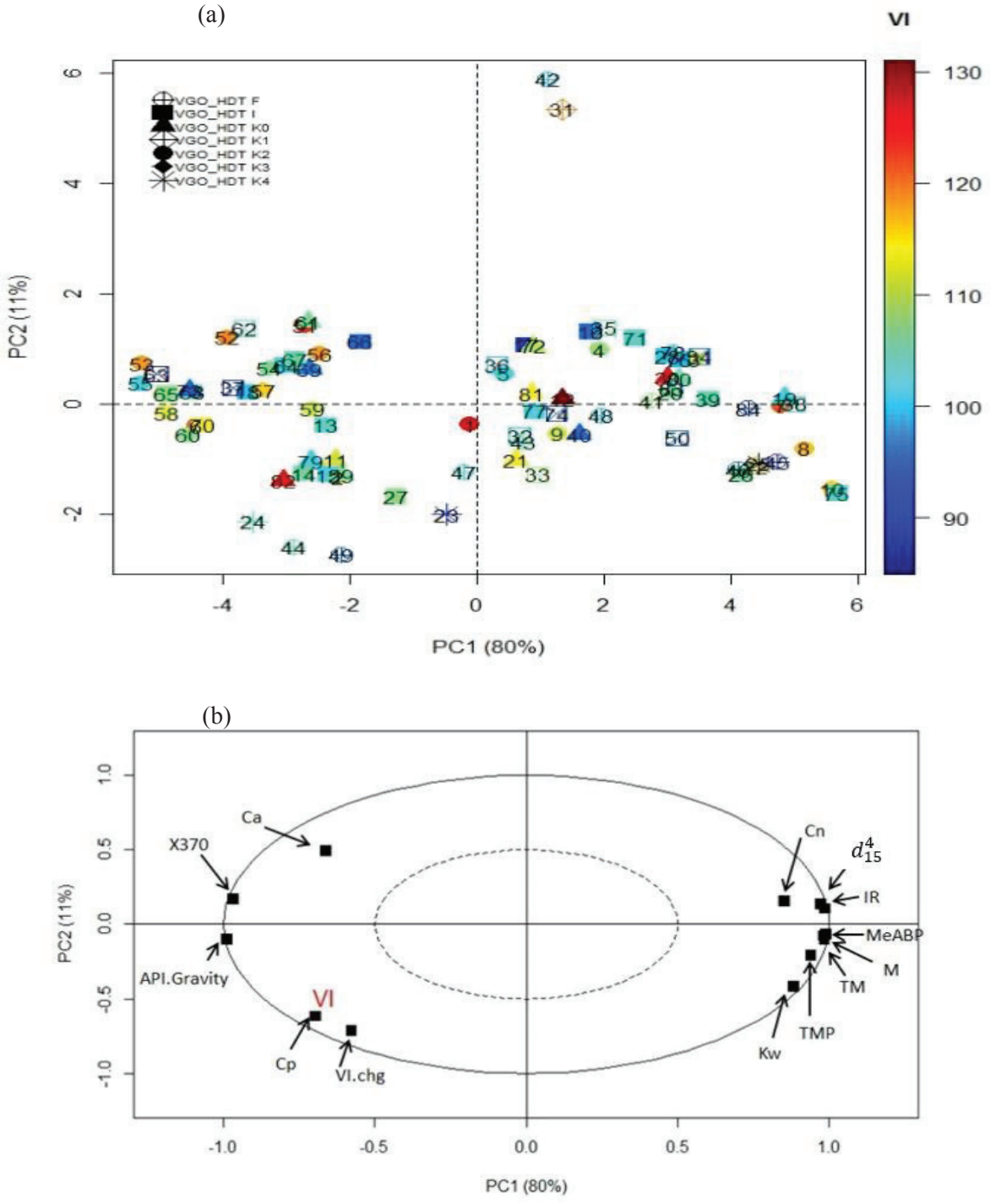


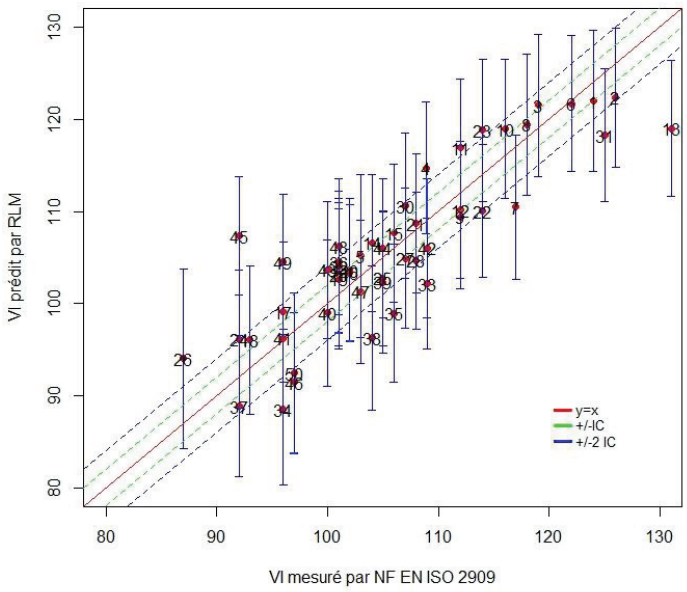
Figure 81 : a) Scores des bilans de liqtot HCK sur le premier plan factoriel ; b) Cercle de corrélation des variables sur le premier plan factoriel.

A la suite de l'ACP, 7 descripteurs ont été sélectionnés : IR, C_a, C_p, K_w, TM, X_{370} et $VI.chg$. La méthode appliquée pour la sélection de variables est la maximisation du $R^2_{ajusté}$.

8.2.2.2 Comparaison des modèles de prédiction du VI

Dans le cas des huiles issues de l'HCK des DSV, le graphe de parité obtenu sur la base d'apprentissage (50 bilans) semble meilleur pour le modèle de krigeage (Figure 82b) que pour le modèle RLM (Figure 82a). Ce constat est appuyé par les statistiques de performance des modèles qui sont précisées dans le Tableau 36 : on note premièrement des valeurs de RMSE de 3,5 pour le modèle de krigeage contre 4,9 pour le modèle RLM (Tableau 36).

(a)



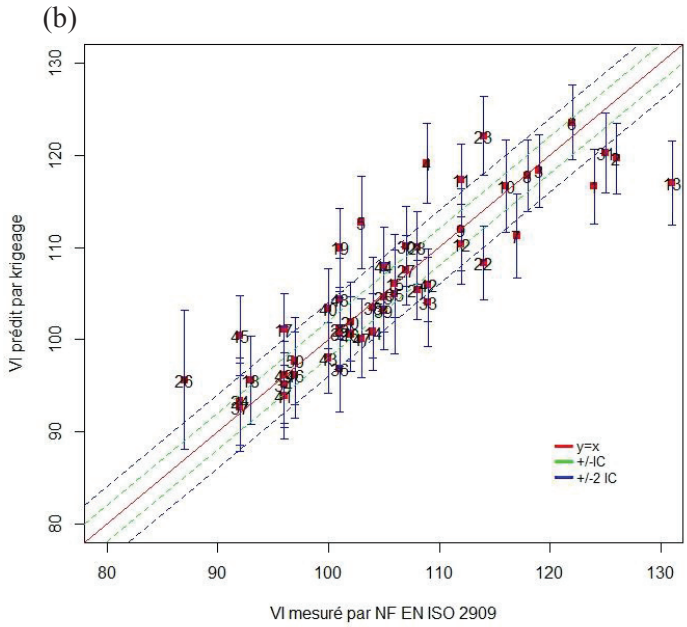


Figure 82 : Graphes de parité des modèles de prédiction du VI de la coupe huile issue de l'HCK des DSV par *leave-one-out* sur la base d'apprentissage. a) modèle RLM ; b) modèle de krigeage

Comme dans le cas du PT de la coupe gazole, les amplitudes variables des intervalles de confiance associés aux prédictions par krigeage permettent de discuter de la position d'un point par rapport au reste des données. Par exemple, l'échantillon 26 est mal prédit mais avec une incertitude relativement forte. La projection de la base d'apprentissage dans l'espace tridimensionnelle (IR, X_{370}, VI.chg) est donnée sur la Figure 83. On observe clairement que l'échantillon 26 (entouré en bleu), est dans cet espace relativement éloigné du reste de la base. Un constat identique peut être fait pour les échantillons 5, 7 ou 23 (zone entourée en vert).

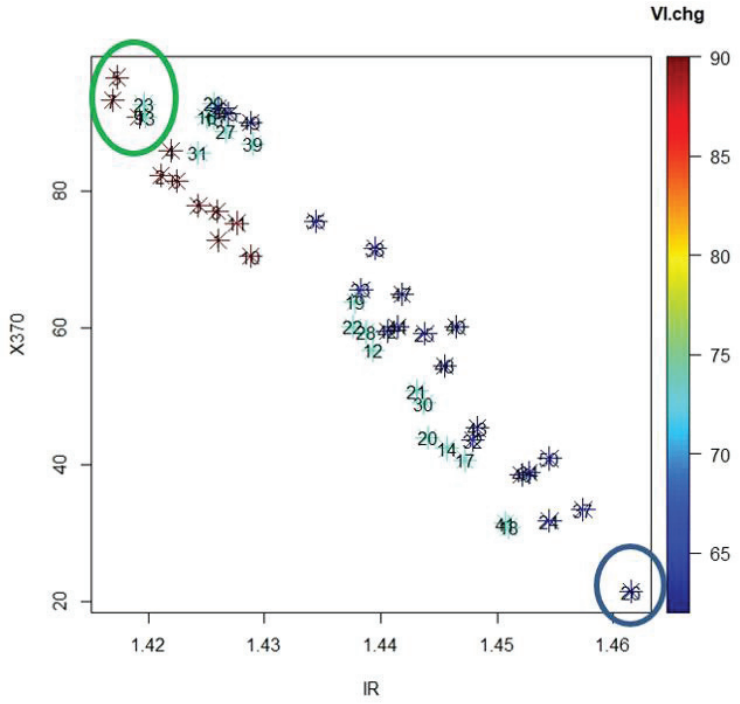


Figure 83 : Projection de la base d'apprentissage dans l'espace (IR, X_{370}, VI.chg)

Sur la base de test (32 bilans), la comparaison des modèles est globalement identique à celle qui a été faite sur la base d'apprentissage. Les graphes de parité correspondants sont représentés sur la Figure 84. Les statistiques des modèles sur la base de test sont clairement meilleures dans le cas du krigeage puisqu'une RMSEP de 2,6 a été obtenue contre 3,6 pour le modèle RLM (Tableau 36). Le krigeage fournit par ailleurs un meilleur pourcentage de points prédits avec une erreur inférieur à l'IC de la mesure (62,5% contre 47%). Ces résultats montrent que le krigeage est plus performant que la RLM dans le cas de la prédiction du VI des huiles issues de l'HCK des DSV. La différence de performances du modèle RLM entre le cas des huiles issues d'HDT et celui des huiles issues d'HCK s'explique par l'impact des catalyseurs utilisés pour ces étapes. Dans le cas de l'HDT, les catalyseurs ont un moindre sur la structure des composés que dans le cas de l'HCK où les réactions d'isomérisation sont très favorisées.

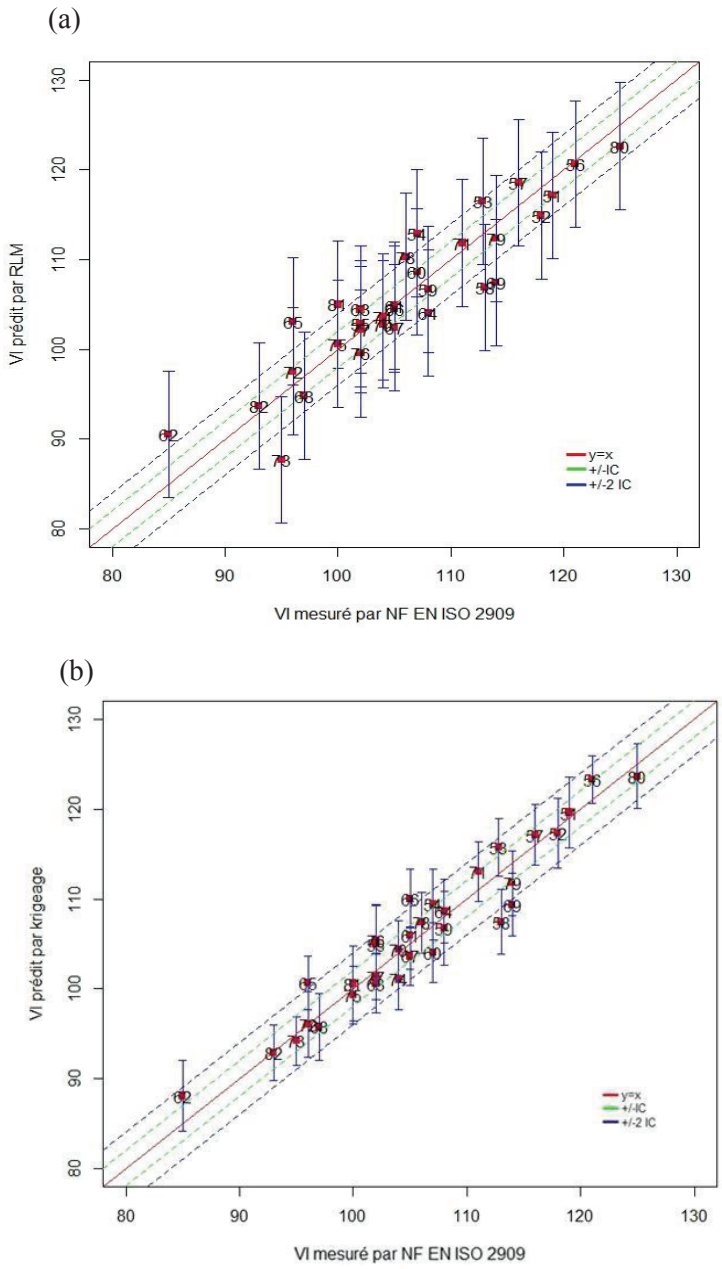


Figure 84 : Graphes de parité des modèles de prédiction du VI de la coupe huile issue de l'HCK des DSV sur la base de test. a) Modèle RLM ; b) Modèle de krigeage

Une comparaison a été effectuée par rapport au modèle IFPEN actuel. La méthode est identique à celle qui a été utilisée dans le cas du PT de la coupe gazole et du VI de la coupe huile produite par HDT. Sur la base d'apprentissage comme sur la base de test, les statistiques montrent que les modèles proposés dans cette étude fournissent de meilleures performances que le modèle IFPEN actuel. Ces résultats valident la méthodologie mise en place en vue d'améliorer la prédiction dans le simulateur IFPEN.

Tableau 36 : Statistiques des modèles de prédiction du VI de la coupe huile issue de l'HCK des DSV à partir de propriétés globales du liqtot

Base d'évaluation	Modèle	MAD	RMSE*	$\tau_{\pm IC}$ (%)	$\tau_{\pm 2IC}$ (%)
Apprentissage (*Leave-one-out*)	RLM	4,0	4,9	36	68
	Krigeage	3,5	4,6	40	74
	Simulateur IFPEN actuel	4,4	6,3	36	68
Test	RLM	2,6	3,6	47	75
	Krigeage	1,8	2,6	62	90
	Simulateur IFPEN actuel	4,2	5,2	28	69

IC →intervalle de confiance de la méthode NF EN ISO 2909 [169] ; * → RMSECV pour la base d'apprentissage et RMSEP pour la base de test.

8.3 Conclusion

Dans notre étude, un modèle RLM et un modèle de krigeage ont été développés à partir des mêmes descripteurs, d'abord pour la prédiction du PT de la coupe gazole, puis pour la prédiction du VI de la coupe huile. Dans le cas du PT de la coupe gazole, les résultats montrent que le krigeage fournit de meilleures performances que le modèle RLM. Dans le cas du VI de la coupe huile, deux cas ont été traités : le premier cas concerne les huiles issues de tests HDT qui ont des VI relativement bas du fait de la quantité d'aromatiques qu'elles contiennent ; le second cas fait référence aux huiles issues d'HCK qui ont naturellement des VI plus élevés. De plus dans chaque cas, les descripteurs sont issus soit de propriétés mesurées sur la charge, soit de propriétés prises sur l'effluent total. Dans le cas des huiles issues de tests HDT, le modèle linéaire et le modèle de krigeage ont des performances comparables. Dans le cas des huiles issues de tests HCK, les performances du krigeage sont meilleures. La différence de performance entre le cas HDT et le cas HCK pour le modèle RLM est principalement due aux effets des catalyseurs qui sont nettement différents suivant le procédé. Globalement, ces résultats obtenus sur le VI de la coupe huile sont particulièrement important puisque qu'aucun descripteur utilisé pour ces modèles n'a été mesuré sur la coupe finale (huile) mais essentiellement sur la charge et l'effluent total. Ainsi, le krigeage permet d'accéder à une caractéristique importante de la coupe huile sans être contraint d'effectuer les étapes de distillation et de déparaffinage préalables à l'obtention de cette coupe, aussi bien dans le cas de l'HDT que dans le cas de l'HCK. Ce résultat induit un gain important de temps d'analyse et de consommation de volume d'échantillon. A notre connaissance, il s'agit de plus des premiers modèles de prédiction du PT de la coupe gazole et du VI de la coupe huile par krigeage.

Conclusions et perspectives

L'objectif de ce travail de thèse était d'améliorer la compréhension moléculaire des propriétés produits pour une meilleure prédiction. Dans cette étude, nous nous sommes focalisés sur le point de trouble de la coupe gazole et l'indice de viscosité de l'huile obtenue lors de l'hydrocraquage de DSV. Dans ce contexte, l'étude bibliographique qui a été effectuée préalablement à nos travaux a permis de relever plusieurs points essentiels. Concernant le PT de la coupe gazole, elle a mis en évidence le lien entre cette propriété et le processus de cristallisation des n-paraffines. Elle a également souligné la tendance des isoparaffines à s'opposer à ce processus de cristallisation. Concernant le VI la coupe huile, les influences négatives des aromatiques et positives des molécules paraffiniques ont été précisées. L'importance de la structure moléculaire des isoparaffines a également été soulignée. Cet état de l'art n'a cependant révélé aucune étude de prédiction des propriétés d'intérêt à partir soit du liqtot, qui permettrait d'apporter une solution à la problématique liée aux EHD ; soit des propriétés globales des coupes pour l'amélioration des performances du simulateur IFPEN.

Dans notre étude, nous avons proposé une méthodologie basée d'une part sur la caractérisation moléculaire des échantillons de produits pétroliers, et d'autre part sur l'application de méthodes multivariées afin de répondre aux trois problématiques identifiées : (1) une meilleure compréhension moléculaire du PT de la coupe gazole et du VI de la coupe huile ; (2) une amélioration de la prédiction de ces propriétés dans le simulateur IFPEN ; (3) une accessibilité aux propriétés d'intérêt lors de l'utilisation d'unité EHD. L'application de la régression multivariée parcimonieuse (*sparse* PLS) aux données GC×GC et ^{13}C RMN a permis de retrouver des résultats en accord avec l'approche empirique appliquée pour le PT par GC×GC et le VI par ^{13}C RMN et des conclusions des études de référence. La qualité des résultats de prédiction obtenus montre de plus qu'il est possible de **proposer des modèles de prédiction de qualités produits à partir de très faibles volumes d'effluent pour répondre à la problématique liée à l'utilisation d'EHD**.

Concernant l'amélioration des performances du simulateur IFPEN, l'utilisation du krigeage pour la prédiction de propriétés de produits pétroliers est tout à fait novatrice dans ce domaine. Aussi bien dans le cas du PT de la coupe gazole que dans celui du VI de la coupe huile, la stratégie préconisée fournit une nette amélioration des performances en comparaison du modèle du simulateur IFPEN actuel. Elle fournit également une mesure semi-quantitative de la distance à la base d'apprentissage. Ceci permet de donner une confiance sur la qualité de la prédiction. Dans le cas du VI, **les modèles développés permettent en plus d'accéder à cette propriété lors de l'utilisation d'unités EHD, plus aisément que *via* des données chromatographiques ou spectroscopiques. Pour chaque test EHD, le VI de la coupe huile pourra donc être prédit à partir de données d'entrée de l'effluent total issu du réacteur ou de la charge. Ce résultat constitue une avancée majeure pour le screening de catalyseur d'HCK au sein d'IFPEN.**

En ce qui concerne la suite de ces travaux, les modèles de prédiction du PT à partir de données GC×GC et ^{13}C RMN du liqtot *via* la PLS ou la *sparse* PLS devront être développés puis testés pour confirmer leur faisabilité.

L'identification de marqueurs moléculaires est une problématique de plus en plus utilisée dans le domaine pétrolier. A ce titre, l'utilisation de techniques multivariées parcimonieuses telles que la *sparse* PLS pour la sélection de variable en très grande dimension ouvre la voie pour l'exploitation d'autres techniques de caractérisation de produits. La RMN est une technique particulièrement performante qui peut permettre d'aller plus loin dans la compréhension moléculaire des propriétés complexes. Des variantes telles que la RMN 2D ou la RMN en motifs structuraux sont des pistes à envisager dans cette optique.

L'utilisation de méthodes d'interpolation (type krigeage) offre plus de flexibilité notamment pour des modèles non linéaires. Ces méthodes permettent en plus de réaliser un diagnostic *a posteriori* des échantillons analysés puisqu'elle fournit pour chaque prédiction une estimation de son incertitude qui dépend de la configuration des données. L'extension du krigeage à la prédiction d'autres propriétés complexes de produits pétroliers est une perspective hautement envisageable malgré une volonté de maintenir un certain niveau d'interprétabilité des modèles. Enfin, il pourrait être intéressant de coupler le krigeage à des techniques d'analyse multivariée pour étendre son utilisation à des données spectroscopiques ou chromatographiques.

Bibliographie

[1] Agence internationale de l'énergie. World energy outlook 2013. Paris: International energy agency; OECD; 2013.

[2] Ward JW. Hydrocracking processes and catalysts. Fuel Processing Technology 1993;35(1-2):55–85.

[3] Wauquier JP. Le raffinage du pétrole: Pétrole brut, produits pétroliers, schémas de fabrication. Paris: Editions Technip; 1994.

[4] ASTM D86. Standard Test Method for Distillation of Petroleum Products at Atmospheric Pressure; 2004.

[5] Rana MS, Samano V, Ancheyta J, Diaz JAI. A review of recent advances on process technologies for upgrading of heavy oils and residua. Fuel 2007;86(9):1216–31.

[6] Scherzer J, Gruia AJ. Hydrocracking science and technology. New York: Marcel Dekker; 1996.

[7] Becker PJ, Celse B, Guillaume D, Costa V, Bertier L, Guillon E et al. A continuous lumping model for hydrocracking on a zeolite catalysts: Model development and parameter identification. Fuel 2016;164:73–82.

[8] Ling H, Wang Q, Shen B-X. Hydroisomerization and hydrocracking of hydrocracker bottom for producing lube base oil. Fuel Processing Technology 2009;90(4):531–5.

[9] Olschar M, Endisch M, Dimmig T, Kuchling T. Investigation of catalytic hydrocracking of Fischer-Tropsch wax for the production of transportation fuels. Oil Gas-European Magazine 2007;33(4):187–93.

[10] Toulhoat H, Raybaud P. Catalysis by transition metal sulphides: From molecular theory to industrial application. Paris: Editions Technip; 2013.

[11] Celse B, Da Costa J-J, Costa V. Experimental Design in Nonlinear Case Applied to Hydrocracking Model: How Many Points Do We Need and Which Ones? Int. J. Chem. Kinet. 2016;48(11):660–70.

[12] Dulot H, Gonzàles Penas H. Modèle d'hydrocraquage haute pression V1.2: Rapport IFP N°59831; 2007.

[13] Perat C, Lopez-Garcia C, Feugnet F, Dulot H. Etablissement de corrélations du procédé mild HCK: propriétés des effluents: Rapport IFPEN N°59596; 2006.

[14] Chainet F, Rivallan M, Vidalie M. Etude de faisabilité pour la compréhension moléculaire et la prédiction de l'indice de viscosité par des techniques spectrocopiques (IR et RMN du 13C): NT IFPEN 203-13 2013.

[15] Lacoue-Nègre M. Prédiction de l'indice de viscosité de l'huile déparaffinée à partir du spectre RMN du 13C de l'effluent total issu du réacteur: NT IFPEN 077-15; 2015.

[16] NF EN 590. Carburants pour automobiles - Carburants pour moteur diesel (gazole) - Exigences et méthodes d'essai; 2014.

[17] NF EN ISO 12185. Pétroles bruts et produits pétroliers - Détermination de la masse volumique; 1996.

[18] NF EN ISO 23015. Produits Pétroliers - Détermination du Point de Trouble; 1994.

[19] NF ISO 2909. Produits pétroliers - Calcul de l'indice de viscosité à partir de la viscosité cinématique; 2000.

[20] NF EN ISO 4259. Produits Pétroliers - Détermination et application des valeurs de fidélité relatives aux méthodes d'essai; 2006.

[21] NF T60-105. Produits Pétroliers - Détermination du point d'écoulement; 1996.

[22] ASTM D3238. Standard Test Method for Calculation of Carbon Distribution and Structural Group Analysis of Petroleum Oils by the n-d-M Method; 2013.
[23] ASTM D341. Standard Practice for Viscosity-Temperature Charts for Liquid Petroleum Products; 2014.
[24] ASTM D4304. Standard Specification for Mineral and Synthetic Lubricating Oil Used in Steam or Gas Turbines;13; 2012.
[25] ASTM D6371. Standard Test Method for Cold Filter Plugging Point of Diesel and Heating Fuels;05; 2010.
[26] ASTM D7042. Standard Test Method for Dynamic Viscosity and Density of Liquids by Stabinger Viscometer (and the Calculation of Kinematic Viscosity);14; 2014.
[27] ASTM D97. Test Method for Pour Point of Petroleum Products; 2012.
[28] Balabin RM, Safieva RZ. Near-infrared (NIR) spectroscopy for biodiesel analysis: Fractional composition, iodine value, and cold filter plugging point from one vibrational spectrum. Energy and Fuels 2011;25(5):2373–82.
[29] Baptista P, Felizardo P, Menezes JC, Neiva Correia MJ. Multivariate near infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40 °C and density at 15 °C of biodiesel. Talanta 2008;77(1):144–51.
[30] Laxlade J. Analyse des produits lourds du pétrole par spectroscopie infrarouge [Thèse de doctorat]. Lille : Université de Lille 1, 2013.
[31] Reboucas MV, Neto BB de. Near infrared spectroscopic prediction of physical properties of aromatics-rich hydrocarbon mixtures. Journal of Near Infrared Spectroscopy 2001;9(4):263–73.
[32] Adhvaryu A, Perez JM, Duda LJ. Quantitative NMR Spectroscopy for the Prediction of Base Oil Properties. Tribology Transactions 2000;43(2):245–50.
[33] Giraudeau P, Baguet E. Improvement of the inverse-gated-decoupling sequence for a faster quantitative analysis of various samples by C-13 NMR spectroscopy. Journal of magnetic resonance 2006;180(1):110–7.
[34] Edwards JC. A Review of Applications of NMR Spectrocopy in the Petroleum Industry. Spectroscopic Analysis of Petroleum and Lubricants 2011.
[35] Bertoncini F, Courtiade-tholance M, Thiébaut D, Jones T. Gas chromatography and 2D-gas chromatography for petroleum industry: The race for selectivity. Paris, France: Editions Technip; 2013.
[36] Dutriez Thomas. Chromatographie multidimensionnelle vers une caractérisation moléculaire étendue des charges type distillat sous vide et la compréhension de leur réactivité à l'hydrotraitement. Paris, France: Université Pierre et Marie Curie; 2010.
[37] Guiochon G, Marchetti N, Mriziq K, Shalliker RA. Implementations of two-dimensional liquid chromatography. Journal of Chromatography A 2008;1189(1-2):109–68.
[38] Lindsay S, Kealey D. High performance liquid chromatography. United States: John Wiley and Sons,New York, NY; 1987.
[39] Shellie RA, Haddad PR. Comprehensive two-dimensional liquid chromatography. Analytical and Bioanalytical Chemestry 2006;386(3):405–15.
[40] Barnett V, Lewis T. Outliers in statistical data. 2nd ed. Chichester: Wiley; 1984.
[41] Chatterjee S, Hadi AS. Influential Observations, High Leverage Points, and Outliers in Linear Regression. Statistical Science 1986;1(3):415–6.
[42] Xie L, Lu C, Wu X-L. Marine bacterial chemoresponse to a stepwise chemoattractant stimulus. Biophysical journal 2015;108(3):766–74.
[43] Zhang L, Li K. Forward and backward least angle regression for nonlinear system identification. Automatica 2015;53:94–102.

[44] Matsumoto R, Hayashi K, Utsunomiya H. Experimental and numerical analysis of friction in high aspect ratio combined forward-backward extrusion with retreat and advance pulse ram motion on a servo press. Journal of Materials Processing Technology 2014;214(4):936–44.
[45] Cruzeiro AB, Shamarova E. Navier–Stokes equations and forward–backward SDEs on the group of diffeomorphisms of a torus. Stochastic Processes and their Applications 2009;119(12):4034–60.
[46] Azaïs J-M, Bardet J-M. Le modèle linéaire par l'exemple: Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS. 2nd ed. Paris: Dunod; 2012.
[47] Mallows CL. Some Comments on Cp. Technometrics 1973;15(4):661–75.
[48] Burch KJ, Whitehead EG. Melting-point models of alkanes 2004;49:858–63.
[49] Tibshirani R. Regression shrinkage and selection via the lasso. Wiley for The Royal Statistical Society. 1996.
[50] Saltelli A, Chan K, Scott EM. Sensitivity analysis 2000.
[51] Sobol' IM. Sensitivity Estimates for Nonlinear Mathematical Models. MMCE 1993;1:407–14.
[52] Tenenhaus M. La régression PLS: Théorie et pratique. Editions Technip 1998.
[53] Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. PLS Methods 2001;58(2):109–30.
[54] Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B, Statistical methodology 2010;72(1):3–25.
[55] Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 2011;12(1):1–17.
[56] Bard Y. Nonlinear Parameter Estimation. Academic Pr. 1973.
[57] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman. New York: Springer; 2001.
[58] Arnaud M, Emery X. Estimation et interpolation spatiale: Méthodes déterministes et méthodes géostatiques. Paris: Hermes Science; 2000.
[59] Ahlberg JH, Nilson EN, Walsh JL. The Theory of Splines and Their Applications. New York: Academic Press; 1967.
[60] Wahba G, Wendelbeger J. Some New mathematical -Methods for Variational Objective Analysis Using Splines and Cross Validation. Monthly Weather Review 1980;108(8):1122–43.
[61] Cressie NAC. The origins of kriging. Mathematical geology 1990;22(3):239–52.
[62] Goovaerts P. Geostatics for natural Resources evaluation. New York: Oxford University Press; 1997.
[63] McKenna AM, Purcell JM, Rodgers RP, Marshall AG. Heavy Petroleum Composition. 1. Exhaustive Compositional Analysis of Athabasca Bitumen HVGO Distillates by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: A Definitive Test of the Boduszynski Model. Energy and Fuels 2010;24:2929–38.
[64] Valavarasu G, Bhaskar M, Balaraman KS. Mild Hydrocracking—A Review of the Process, Catalysts, Reactions, Kinetics, and Advantages. Petroleum Science and Technology 2003;21(7-8):1185–205.
[65] Wise JJ, Katzer JR, Chen NY. Catalytic Dewaxing in Petroleum Processing. Abstracts of papers of the American Chemical Society 1986;191.
[66] Lynch TR. Process chemistry of lubricant base stocks. Boca Raton: CRC Press; 2008.
[67] Krishna R, Joshi GC, Purohit RC, Agrawal KM, Verma PS, Bhattacharjee S. Correlation of pour point of gas oil and vacuum gas oil fractions with compositional parameters. Energy Fuels 1989;3(1):15–20.

[68] Claudy P, Létoffé J-M, Neff B, Damin B. Diesel fuels: Determination of onset crystallization temperature, pour point and filter plugging point by differential scanning calorimetry. Correlation with standard test methods. Fuel 1986;65(6):861–4.
[69] Merdrignac I, Quignard A. Analyse de la structure globale des distillats atmosphériques: Rapport IFPEN N° 41925; 1995.
[70] Quignard A. Analyse de la structure globale et détaillée des distillats atmosphériques Application au calcul des caractéristiques pétrolières Modélisation des propriétés: Rapport IFPEN N°54043; 2000.
[71] Sastry MI, Chopra A, Sarpal AS, Jain SK, Srivastava SP, Bhatnagar AK. Determination of Physicochemical Properties and Carbon-Type Analysis of Base Oils Using Mid-IR Spectroscopy and Partial Least-Squares Regression Analysis. Energy Fuels 1998;12(5):304–11.
[72] Sarpal AS, Kapur GS, Chopra A, Jain SK, Srivastava SP, Bhatnagar AK. Hydrocarbon characterization of hydrocracked base stocks by one- and two-dimensional NMR spectroscopy. Fuel 1996;75(4):483–90.
[73] Verdier S, Coutinho AP, Silva MS, Alkilde OF, Hansen Ja. A critical approach to viscosity index. Fuel 2009, 2009:2199–206.
[74] Besse P, Béatrice L. Apprentissage Statistique: Modélisation, prévision, data mining; Available from: https://www.math.univ-toulouse.fr/~besse/pub/Appren_stat.pdf.
[75] Sammut C, Webb GI (eds.). Encyclopedia of Machine Learning. Boston, MA: Springer US; 2010.
[76] Vapnik VN. An overview of statistical learning theory. IEEE Transactions on Neural Networks 1999;10(5):988–99.
[77] Haykin SS. Neural networks: A comprehensive foundation. New York, Toronto, New York: Second Edition; 1994.
[78] Guermeur Y, Elisseeff A, Paugam-Moisy H. Estimating the sample complexity of a multi-class discriminant model. Artificial Neural Network 1999, 1999:310–5.
[79] Schölkopf B, Smola AJ. Learning with kernels: Support vector machines, regularization, optimization, and beyond. Cambridge, Mass.: MIT Press; 2002.
[80] Tsang CY, Ker VSF, Miranda RD, Wesch JC. Equation predicts Diesel Cloud Points. Oil & Gas Journal 1988;86(13).
[81] Products P. Standard Test Method for Cloud Point of Petroleum Products; 2011.
[82] NF EN 116. Produits Pétroliers - Détermination de la Température Limite de Filtrabilité; 1997.
[83] Daage M. Zeolites for cleaner technologies 2002;3:167–87.
[84] Denis J, Durand JP. Modification of Wax Crystallization in Petroleum Products. Revue de L'Institut Français du Pétrole 1991;46(5):637–49.
[85] Petinelli JC. Effect of ninyl ethylene acetate copolymers on the nucleation and growth kinetics of n-paraffins in a hydrocarbon medium. Revue de L'Institut Français du Pétrole 1979;34(5):791–811.
[86] Petinelli JC. Influence of additives on the crystallization of n-paraffins in a hydrocarbon medium. Revue de L'Institut Français du Pétrole 1979;34(5):771–90.
[87] Turner WR. Normal Alkanes. Industrial & Engineering Chemistry Product Research and Development 1971;10(3).
[88] Rossemyr LI. Cold Flow Properties and Response to Cold Flow Improver of Some Typical Fuel Oils. Industrial & Engineering Chemistry Product Research and Development 1979;18(3):227–30.
[89] Hübschmann H-J. Handbook of GC-MS: Fundamentals and applications. Weinheim, New York: Wiley-VCH; 2001.
[90] Riazi MR. Characterization and properties of petroleum fractions. 1st ed. West Conshohocken, PA: ASTM International; 2005.
[91] Reddy SR. A thermodynamic model for predicting n-paraffin crystallization in diesel fuels. Fuel 1986;65(12):1647–52.

[92] Cookson DJ, Latten JL, Shaw IM, Smith BE. Property-composition relationships for diesel and kerosene fuels. Fuel 1985;64(4):509–19.
[93] Cookson DJ, Smith BE. One-dimensional and two-dimensional NMR methods for elucidating structural characteristics of aromatic fractions from petroleum and synthetic fuels. Energy and Fuels 1987;1(1):111–20.
[94] Cookson DJ, Smith BE. Calculation of Jet and Diesel Fuel Properties Using 13C NMR Spectroscopy. Energy Fuels 1990;4(6):152–6.
[95] Cookson DJ, Iliopoulos P, Smith BE. Composition-property relations for jet and diesel fuels of variable boiling range. Fuel 1995;74(1):70–8.
[96] Souchon V, Caillol N. Rapport IFPEN NT-148-14: Modélisation de propriétés dans les coupes gazoles à partir des données GC×GC-FID; 2014.
[97] Dulot H, Gonzàles Penas H. Rapport IFP N°59831: Modèle d'hydrocraquage haute pression V1.2; 2007.
[98] ASTM D2887. Test Method for Boiling Range Distribution of Petroleum Fractions by Gas Chromatography: ASTM International; 2016.
[99] Mendez A, Bruzual J. Molecular Characterization of Petroleum and Its Fractions By Mass Spectrometry. In: Hsu CS, editor. Analytical Advances for Hydrocarbon Research. Boston, MA: Springer US; 2003, p. 73–93.
[100] Souchon V, Caillol N. Modélisation de propriétés dans les coupes gazoles à partir des données GC×GC-FID: NT IFPEN 148-14; 2014.
[101] Marinovic S, Bolanca T, Ukic S. Prediction of Diesel Fuel Cold Properties Using Artificial 2012;48(1):47–51.
[102] Pasadakis N, Sourligas S, Foteinopoulos C. Prediction of the distillation profile and cold properties of diesel fuels using mid-IR spectroscopy and neural networks. Fuel 2006;85(7-8):1131–7.
[103] Wu C, Zhang J, Li W, Wang Y, Cao H. Artificial neural network model to predict cold filter plugging point of blended diesel fuels. Fuel Processing Technology 2006;87(7):585–90.
[104] ASTM D2270. Standard Practice for Calculating Viscosity Index from Kinematic Viscosity at 40 and 100°C;(95); 1991.
[105] Institute TAP, Technology APIDoS. Properties of Hydrocarbons of High Molecular Weight. American Petroleum Institute Division of Science and Technology 1967.
[106] Briant J, Denis J, Parc G. Rheological Properties of Lubricants. Paris: Éditions Technip; 1989.
[107] Kapur GS, Oil BI. Temperature-Dependent Effects in Base Oils: Carbon-I3 NMR Spin-Lattice Relaxation Time and Viscometry Studies. Lubr. Sci. 2002;00(14):287–302.
[108] Sarpal AS, Kapur GS, Mukherjee S, Jain SK. Characterization by 13C NMR spectroscopy of base oils produced by different processes. Fuel 1997;76(10):931–7.
[109] Sarpal AS, Sastry MI, Bansal V, Singh I, Mazumdar SK, Basu B. Correlation of structure and properties of groups I to III base oils. Lubr. Sci. 2012:199–215.
[110] Abdel-Rahman EM, Mutanga O, Odindi J, Adam E, Odindo A, Ismail R. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. Computers and Electronics in Agriculture 2014;106(0):11–9.
[111] Al-Ghouti MA, Al-Degs YS, Amer M. Application of chemometrics and FTIR for determination of viscosity index and base number of motor oils. Talanta 2010;81(3):1096–101.
[112] Sastry MIS, Chopra A, Sarpal AS, Srivastava SP, Bhatnagar AK. Carbon type analysis of hydrotreated and conventional lube-oil base stocks by IR spectroscopy 1996;15(12):1471–5.

[113] Sarpal AS, Kapur GS, Chopra A, Jain SK, Srivastava SP, Bhatnagar AK. Hydrocarbon characterization of hydrocracked base stocks by one- and two-dimensional NMR spectroscopy. Fuel 1996;75(4):483–90.
[114] Sperber O, Kaminsky W, Geissler A. Structure analysis of paraffin waxes by 13C-NMR spectroscopy. Petroleum Science and Technology 2005;23(1):47–54.
[115] Chainet F. Compréhension et modélisation des propriétés: NT IFPEN 034-14; 2014.
[116] Kobayashi M, Saitoh M, Ishida K, Yachi H. Viscosity properties and molecular structure of lube base oil prepared from Fischer-Tropsch waxes. Journal of the Japan Petroleum Institute 2005;48(6):365–72.
[117] Sharma BK, Adhvaryu A, Perez JM, Erhan SZ. Effects of hydroprocessing on structure and properties of base oils using NMR. Fuel Processing Technology 2008;89(10):984–91.
[118] Sharma BK, Adhvaryu A, Sahoo SK, Stipanovic AJ, Erhan SZ. Influence of Chemical Structures on Low-Temperature Rheology Oxidative Stability and Physical Properties of Group II and III Base Oils 2004(2):952–9.
[119] Hennico A, Billon A, Bigeard PH, Peries JP. Ifp's New Flexible Hydrocracking process Combines Maximum Conversion with Production of High-Viscosity, High-VI Lube Stocks. Revue de L'Institut Français du Pétrole 1993;48(2):127–39.
[120] Braga JW, Santos AA, Martins IS. Determination of viscosity index in lubricant oils by infrared spectroscopy and PLSR. Fuel 2014;120:171–8.
[121] Consden R, Gordon AH, Martin AJ. Qualitative analysis of proteins: A partition chromatographic method using paper. The Biochemical journal 1944;38(3):224–32.
[122] Giddings JC. Two-dimensional separations - Concept and promise. Anal. Chem. 1984;56(12):1258-&.
[123] Koning S, Janssen HG, Brinkman UAT. Group-type characterisation of mineral oil samples by two-dimensional comprehensive normal-phase liquid chromatography-gas chromatography with time-of-flight mass spectrometric detection. Journal of Chromatography A 2004;1058(1-2):217–21.
[124] Koning S, Janssen HG, van Deursen M, Brinkman UAT. Automated on-line comprehensive two-dimensional LC x GC and LC x GC-ToF MS: Instrument design and application to edible oil and fat analysis. Journal of Separation Science 2004;27(5-6):397–409.
[125] Venter A, Rohwer ER. Comprehensive two-dimensional supercritical fluid and gas chromatography with independent fast programmed heating of the gas chromatographic column. Anal. Chem. 2004;76(13):3699–706.
[126] Liu ZY, Phillips JB. Comprehensive 2-dimesional gas -chromatography using an on-column thermal modulator interface. Journal of chromatographic science 1991;29(6):227–31.
[127] Adahchour M, Beens J, Brinkman UAT. Recent developments in the application of comprehensive two-dimensional gas chromatography. Journal of chromatography. A 2008;1186(1-2):67–108.
[128] Dorman FL, Whiting JJ, Cochran JW, Gardea-Torresdey J. Gas Chromatography. Anal. Chem. 2010;82(12):4775–85.
[129] Mondello L, Bartle KD, Lewis A. Multidimensional chromatography. Chichester: Wiley; 2001.
[130] Ramos L. Comprehensive two dimensional gas chromatography. Amsterdam: Elsevier; 2009.
[131] Boursier L. Caractérisation et réactivité en hydrotraitement des composés hétéroatomiques présents dans les distillats sous vide du pétrole [Thèse de doctorat de Chimie Analytique]: Université Pierre et Marie Curie, 2014.
[132] Dutriez T. Chromatographie multidimensionnelle vers une caractérisation moléculaire étendue des charges type distillat sous vide et la compréhension de leur réactivité à l'hydrotraitement [Thèse de doctorat]. Paris, France: Université Pierre et Marie Curie, 2010.

[133] Williams RB. Characterization of Hydrocarbons in Petroleum by Nuclear Magnetic Resonance Spectrometry. In: Symposium on Composition of Petroleum Oils, Determination and Evaluation. 100 Barr Harbor Drive, PO Box C700, West Consho hocken, PA 19428-2959: ASTM International; 1958, 168-168-27.
[134] Cookson DJ, Rolls CL, Smith BE. Structural characteristics of branched plus cyclic sturates from petroleum and coal derived diesel fuels. Fuel 1989;68(6):788–92.
[135] Mergui S, Lallemand JY. Application de la R.M.N. à deux dimensions: Détermination de structure de produits naturels analyse de mélanges coupes pétrolières (L.C.O.) applicabilité de la R.M.N. 2D à l'étude des milieux vivants [Thèse de doctorat de Physique]. Université de Paris-Sud; 1989.
[136] Kapur GS, Berger S. Simplification and assignment of proton and two-dimensional hetero-correlated NMR spectra of petroleum fractions using gradient selected editing pulse sequences. Fuel 2002;81(7):883–92.
[137] ASTM 1218. Test Method for Refractive Index and Refractive Dispersion of Hydrocarbon Liquids;16; 2016.
[138] ASTM D445. Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity);15; 2010.
[139] ASTM D5291. Standard test methods for instrumental determination of carbon, hydrogen, and nitrogen in petroleum products and lubricants. ASTM;10; 2010.
[140] Souchon V, Mouillet Y. Distillats sous vide et effluents totaux d'hydrocraquage Analyse et distillation simulée par famille par chromatographie en phase gazeuse: Méthode IFPEN 1602; 2016.
[141] Griffith JF, Winniford WL, Sun K, Edam R, Luong JC. A reversed-flow differential flow modulator for comprehensive two-dimensional gas chromatography. Journal of chromatography. A 2012;1226:116–23.
[142] Savorani F, Tomasi G, Engelsen SB. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. Journal of magnetic resonance 2010;202(2):190–202.
[143] Sousa SAA, Magalhães A, Ferreira MMC. Optimized bucketing for NMR spectra: Three case studies. Chemometrics and Intelligent Laboratory Systems 2013;122:93–102.
[144] Parlov Vuković J, Novak P, Plavec J, Friedrich M, Marinić Pajc L, Hrenar T. NMR and Chemometric Characterization of Vacuum Residues and Vacuum Gas Oils from Crude Oils of Different Origin. Croat. Chem. Acta 2015;88(1):89–95.
[145] Peinder P, Visser T, Petrauskas DD, Salvatori F, Soulimani F, Weckhuysen BM. Partial least squares modeling of combined infrared, 1H NMR and 13C NMR spectra to predict long residue properties of crude oils. Vibrational Spectroscopy 2009;51(2):205–12.
[146] Trbovic N, Dancea F, Langer T, Günther U. Using wavelet de-noised spectra in NMR screening. Journal of magnetic resonance 2005;173(2):280–7.
[147] Jellema R. Variable Shift and Alignment. Comprehensive Chemometrics 2009;2.
[148] Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 1964;36(8):1627–39.
[149] Jolliffe IT. Principal component analysis. 2nd ed. Berlin, London: Springer; 2002.
[150] Benzecri J-P. Histoire et prehistoire de l'analyse des donnees. Paris: Dunod; 1982.
[151] Volle M. Analyse des données. 4th ed. Paris: Economica; 1997.
[152] Celeux G. Classification automatique des données. Paris: Dunod; 1989.
[153] Wold S. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 1987;2:37–52.
[154] Cormack RM. A review of classification. [S.l.: s.n.]; 1971.

[155] Szekely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. Journal of Classification 2005;22(2):151–83.
[156] M. Furnival G, W. Wilson R. Regression by Leaps and Bounds. Technometrics 1974;16.
[157] Bach F. Bolasso: Model consistent Lasso estimation through the bootstrap 2008:33–40.
[158] Cukier RI, Fortuin CM, Shuler KE, Petschek AG, Schaibly JH. A Study of the Sensitivity of Coupled Reaction Systems to Uncertainties in Rate Coefficients. I. Theory. Ft. Belvoir: Defense Technical Information Center; 1973.
[159] McKay MD. Evaluating prediction uncertainty. Washington, DC: The Commission; 1995.
[160] Martens H. Multivariate calibration. Chichester: J. Wiley; 2002.
[161] Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 2011;12(1):1–17.
[162] Stein M. Gaussian approximations to conditional distributions for multi-Gaussian processes. Mathematical geology 1987;19(5):387–405.
[163] Myung IJ. Tutorial on maximum likelihood estimation. Journal of Mathematical Psychology 2003;47(1):90–100.
[164] Chaudhuri P, Mykland PA. Nonlinear Experiments: Optimal Design and Inference Based on Likelihood 1993;88(422):538.
[165] Cressie NAC. Statistics for spatial data. New York, Chichester: Wiley; 1993.
[166] Dubrule O. Comparing splines and kriging. Computers and Geosciences 1984;10(2-3):327–38.
[167] Matheron G. Splines and Kriging: Their Formal Equivalence. Fontainebleau; 1975.
[168] NF EN ISO 23015. Produits Pétroliers - Détermination du Point de Trouble; 1994.
[169] NF ISO 2909. Produits pétroliers - Calcul de l'indice de viscosité à partir de la viscosité cinématique;75.080; 2000.
[170] Sperber O, Kaminsky W, Geissler A. Structure analysis of paraffin waxes by C-13-NMR spectroscopy. Petroleum Science and Technology 2005;23(1):47–54.
[171] Sastry MI, Chopra A, Sarpal AS, Jain SK, Srivastava SP, Bhatnagar AK. Determination of Physicochemical Properties and Carbon-Type Analysis of Base Oils Using Mid-IR Spectroscopy and Partial Least-Squares Regression Analysis. Energy Fuels 1998;12(5):304–11.
[172] Sarpal AS, Kapur GS, Chopra A, Jain SK, Srivastava SP, Bhatnagar AK. Hydrocarbon characterization of hydrocracked base stocks by one- and two-dimensional n.m.r. spectroscopy. Fuel 1996;75(4):483–90.
[173] ASTM D6300. Standard Practice for Determination of Precision and Bias Data for Use in Test Methods for Petroleum Products and Lubricants;15; 2015.
[174] Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 2000;50(1):1–18.
[175] Da Costa Soares J-J. Compréhension moléculaire pour la prédiction propriétés physico-chimiques dans les produits pétroliers Rapport Bibliographique: NT IFPEN L061; 2016.
[176] Jouan-Rimbaud D, Bouveresse E, Massart DL, Noord OE de. Detection of prediction outliers and inliers in multivariate calibration. Analytica chimica acta 1999;388(3):283–301.
[177] Walczak B, Massart DL. Robust Principal Components Regression as A Dectection Tool For Outliers. Chemometrics and Intelligent Laboratory Systems 1995;27(1):41–54.

Annexe

Annexe A : Exemples de composés présents dans les pétroles bruts

Tableau A. 1 : Exemples de composés soufrés présents dans les pétroles bruts [30]

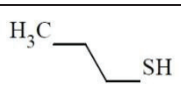
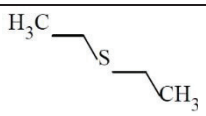
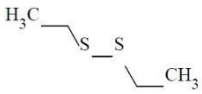
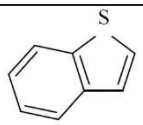
Familles	mercaptans	sulfures	disulfures	benzothiophènes
Exemples				

Tableau A. 2: Exemples de composés azotés présents dans les pétroles bruts [30]

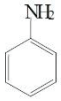
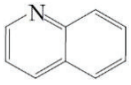
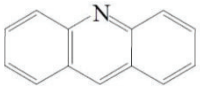
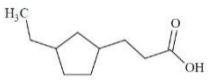
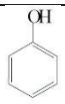
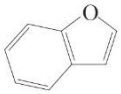
	Dérivés basiques				**Dérivés neutres**	
Familles	Aniline	Pyridine	Quinoléine	Acridine	Pyrrole	Indole
Exemples						

Tableau A. 3 : Exemples de composés oxygénés présents dans les pétroles bruts [30]

Familles	acide naphténique	phénol	eurane	benzofurane
Exemples				

Annexe B : Exemples de réactions d'hydrotraitement et d'hydrocraquage

Tableau A. 4 : Exemples de réactions d'hydrotraitement [10]

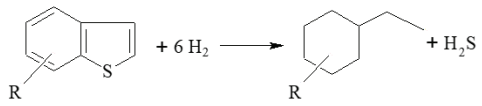
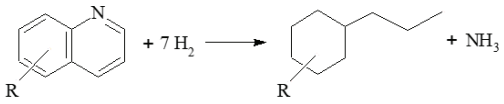
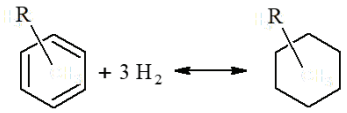
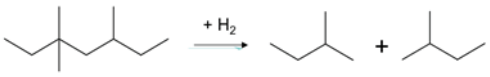
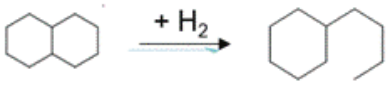
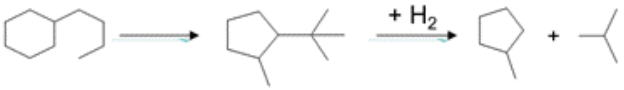
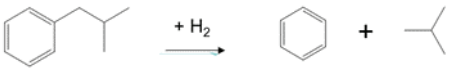
Réaction	HDS	
	HDN	
	HDA	

Tableau A. 5 : Exemples de réactions d'hydrocraquage [10]

Paraffines	Isomérisation	
	Craquage	
Naphtènes	Ouverture de cycle	
	Isomérisation + craquage	
Aromatiques	Désalkylation	

Annexe C: Exemple de calcul du VI et fidélités de la méthode

La procédure de calcul de l'indice de viscosité est la suivante [19] :

- Si U>H (*i.e.* VI<100)
  - $VI = \frac{L-H}{L-U} \times 100$
- Si U<H (*i.e.* VI>100)
  - $VI = \frac{(10^N - 1)}{0{,}00715} + 100$ avec $N = \frac{\log H - \log U}{\log Y}$

Où :

- U est la viscosité cinématique (mm²/s) à 40°C d'un produit pétrolier dont le VI est à mesurer ;
- L est la viscosité cinématique (mm²/s) à 40°C d'un produit pétrolier dont l'indice de viscosité est 0 et ayant la même viscosité cinématique à 100°C que le produit pétrolier dont l'indice de viscosité est à mesurer ;
- H est la viscosité cinématique (mm²/s) à 40°C d'un produit pétrolier dont l'indice de viscosité est 100 et ayant la même viscosité cinématique à 100°C que le produit pétrolier dont l'indice de viscosité est à mesurer ;
- Y est la viscosité cinématique (mm²/s) à 100°C du produit pétrolier dont l'indice de viscosité doit être déterminé.

Les valeurs de L et H sont données dans les normes pour des viscosités cinématiques à 100°C comprises entre 2 et 70 mm²/s.

Exemple :

Soit une huile dont les caractéristiques déterminées par la mesure sont les suivantes :

- $\nu_{40} = U = 73{,}30\ mm^2/s$ et $\nu_{100} = 8{,}86\ mm^2/s$

Les valeurs de H et L sont déterminées grâce à des tableaux (disponibles dans la norme NF ISO 2909 [169]) par interpolation :

- L=119,94 et H=69,48

On en déduit le VI suivant le calcul :

$$\boldsymbol{VI = \frac{(119{,}94 - 69{,}48)}{(119{,}94 - 73{,}40)} \times 100 = 92} \qquad \textbf{(Eq. A. 1)}$$

Tableau A. 6 : Normes et fidélités des méthodes de détermination du VI selon NF EN ISO 2909 [12]

Fidélités	r						R						IC					
Viscosité à 100°C (cSt)	4	6	8	15	30	50	4	6	8	15	30	50	4	6	8	15	30	50
Grade	100N	200N	350N	850N														
Mode A	0<VI<100																	
30	0,9	0,6	0,5	0,4	0,3	0,3	5,3	3,6	2,9	2,2	1,8	1,7	3,8	2,6	2,0	1,6	1,3	1,2
50	0,9	0,6	0,4	0,3	0,3	0,2	5,0	3,3	2,6	1,9	1,6	1,4	3,6	2,3	1,8	1,4	1,1	1,0
70	0,8	0,5	0,4	0,3	0,2	0,2	4,8	2,9	2,2	1,6	1,3	1,1	3,4	2,1	1,6	1,2	0,9	0,8
90	0,8	0,4	0,3	0,2	0,2	0,1	4,5	2,5	1,9	1,3	1,0	0,8	3,2	1,8	1,4	0,9	0,7	0,6
100	0,7	0,4	0,3	0,2	0,1	0,1	4,3	2,4	1,8	1,2	0,8	0,7	3,1	1,7	1,2	0,8	0,6	0,5
Mode B	100<VI<200																	
100	0,5	0,4	0,3	0,2	0,2	0,2	2,9	2,2	1,8	1,4	1,1	1,0	2,1	1,5	1,3	1,0	0,8	0,7
110	0,5	0,4	0,3	0,2	0,2	0,2	3,1	2,3	1,9	1,4	1,2	1,0	2,2	1,6	1,4	1,0	0,8	0,7
120	0,6	0,4	0,3	0,3	0,2	0,2	3,3	2,4	2,0	1,5	1,2	1,1	2,3	1,7	1,4	1,1	0,9	0,8
130	0,6	0,4	0,4	0,3	0 ,2	0,2	3,4	2,5	2,1	1,6	1,3	1,1	2,4	1,8	1,5	1,1	0,9	0,8
140	0,6	0,5	0,4	0,3	0,2	0,2	3,6	2,6	2,2	1,7	1,4	1,2	2,5	1,9	1,6	1,2	1,0	0,8

Annexe D : Quelques notions de statistiques mathématiques

Notion d'estimateur

Un estimateur est une statistique (au sens mathématique) qui permet d'évaluer un paramètre inconnu relatif à une loi de probabilité (par exemple son espérance mathématique ou sa variance). Soit θ un paramètre associé à une loi de probabilité d'une variable aléatoire X. On souhaite estimer θ à partir de n réalisations $(x_1, \dots, x_n)$ de X. On appelle estimateur de θ et on note $\hat{\theta}_n$ toute fonction qui dépend uniquement de l'ensemble des valeurs observées. Par exemple, la dite « loi des grand nombres » assure que pour toute variable aléatoire X, la moyenne empirique définie par [46] :

$$\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \textbf{(Eq. A. 2)}$$

est un estimateur de son espérance mathématique.

Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le « meilleur » estimateur, c'est-à-dire celui qui donnera une estimation ponctuelle la plus proche possible du paramètre, et ceci quel que soit les observations dont on dispose. La qualité d'un estimateur est donnée par deux grandeurs distinctes :

- Son biais qui représente la différence entre la valeur réelle du paramètre à estimer et l'espérance mathématique de l'estimateur

$$\boldsymbol{Biais}(\hat{\boldsymbol{\theta}}_n) = \boldsymbol{E}[\hat{\boldsymbol{\theta}}_n] - \boldsymbol{\theta} \qquad \textbf{(Eq. A. 3)}$$

- Sa variance

$$\boldsymbol{V}[\hat{\boldsymbol{\theta}}_n] = \boldsymbol{E}\left[\left(\boldsymbol{E}[\hat{\boldsymbol{\theta}}_n] - \boldsymbol{\theta}\right)^2\right] \qquad \textbf{(Eq. A. 4)}$$

où $E[\hat{\theta}_n]$ représente l'espérance mathématique de l'estimateur $\hat{\theta}_n$. Un estimateur est dit « optimal » s' il est sans biais et de variance minimale. Notez qu'un estimateur dépend naturellement des observations dont on dispose. A ce titre, il peut être considéré comme une variable aléatoire. Ce qui justifie qu'on lui associe une espérance mathématique et une variance.

Fidélités des méthodes de mesures standard

Ce paragraphe rappelle quelques notions de base de statistiques couramment utilisées pour caractériser les méthodes de mesure et les modèles de prédiction. Ces notions permettent de quantifier la précision et la robustesse des différentes méthodes. De plus amples informations sont disponibles dans le rapport n° 59556 [13] et dans les normes qui définissent ces notions, à savoir : les normes ASTM D6300-13 [173] et NF ISO 4259 [20]. Ces statistiques ou fidélités sont liées à des distributions normales et sont définies pour un niveau de risque de 5%.

Notion de répétabilité

La différence entre deux résultats individuels obtenus par le même opérateur, sur une matière identique soumise à essai, par une analyse utilisant le même appareillage dans l'intervalle de temps le plus court, ne dépassera pas la valeur de répétabilité r, en moyenne plus d'une fois sur vingt, lors de l'application normale, et correcte de la méthode [20].

Notion de reproductibilité

La différence entre deux résultats uniques et indépendants, obtenus par différents opérateurs travaillant dans des laboratoires différents sur un même produit, ne dépassera pas la valeur de reproductibilité R, au cours d'une longue série d'essais, en moyenne plus d'une fois sur vingt, lors de l'application normale et correcte de la méthode.

Cette reproductibilité ne peut être déterminée que dans le cadre d'essais circulaires sur des méthodes analytiques relativement stables. Compte tenu du coût de tels essais, ils ne sont pas toujours réalisés. Par contre, toute adoption d'une norme présuppose l'existence de la détermination de cette reproductibilité [20].

Reproductibilité intra laboratoire

La différence entre deux résultats uniques et indépendants, obtenus par différents opérateurs travaillant dans le même laboratoire sur un même produit et un même matériel, ne dépassera pas la valeur de reproductibilité intra laboratoire $R_{intra\ labo}$ au cours d'une longue série d'essais, en moyenne plus d'une fois sur vingt, lors de l'application normale et correcte de la méthode [20].

On notera que cette reproductibilité intra laboratoire n'est parfois définie que parce que la méthode n'existe qu'au sein du laboratoire en question. En général, elle est alors très inférieure à la vraie reproductibilité, car le matériel est toujours le même, et les opérateurs sont toujours peu nombreux et tous formés aux mêmes techniques par les mêmes personnes. La répétabilité ou la reproductibilité sont estimées à partir de l'écart-type selon [20] :

$$\boldsymbol{s = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{n-1}} \qquad \textbf{(Eq. A. 5)}$$

où x_i est la i^{ème} valeur obtenue sur une série de n mesures d'un échantillon, $\bar{x}$ est la valeur moyenne et n le nombre de mesures.

Intervalle de confiance d'une mesure

L'intervalle de confiance à 95% d'un résultat de mesure est l'intervalle de valeurs ayant une probabilité de 95% de contenir la valeur vraie si la méthode est non biaisée. Pour un nombre n de mesures, et en notant X_i la valeur de la mesure i, la valeur X s'exprime en fonction de r et R de la manière suivante [20] :

$$\boldsymbol{X = \sum_{i=1}^{n}\frac{X_i}{n} \pm IC} \qquad \textbf{(Eq. A. 6)}$$

où IC est défini comme suit [20] :

$$IC = \frac{\sqrt{R^2 - r^2 \times (1 - \frac{1}{n})}}{\sqrt{2}} \qquad \textbf{(Eq. A. 7)}$$

Dans la quasi-totalité des cas l'analyse n'est faite qu'une fois. De ce fait on a [20] :

$$IC = \frac{R}{\sqrt{2}} \qquad \textbf{(Eq. A. 8)}$$

Enfin, pour les analyses dont on ne dispose pas de la reproductibilité, l'usage veut que l'on prenne le double de la répétabilité, ce qui donne l'intervalle de confiance suivant [20] :

$$IC = r\sqrt{2} \qquad \textbf{(Eq. A. 9)}$$

Annexe E : Structures identifiables par spectroscopie RMN du ^{13}C

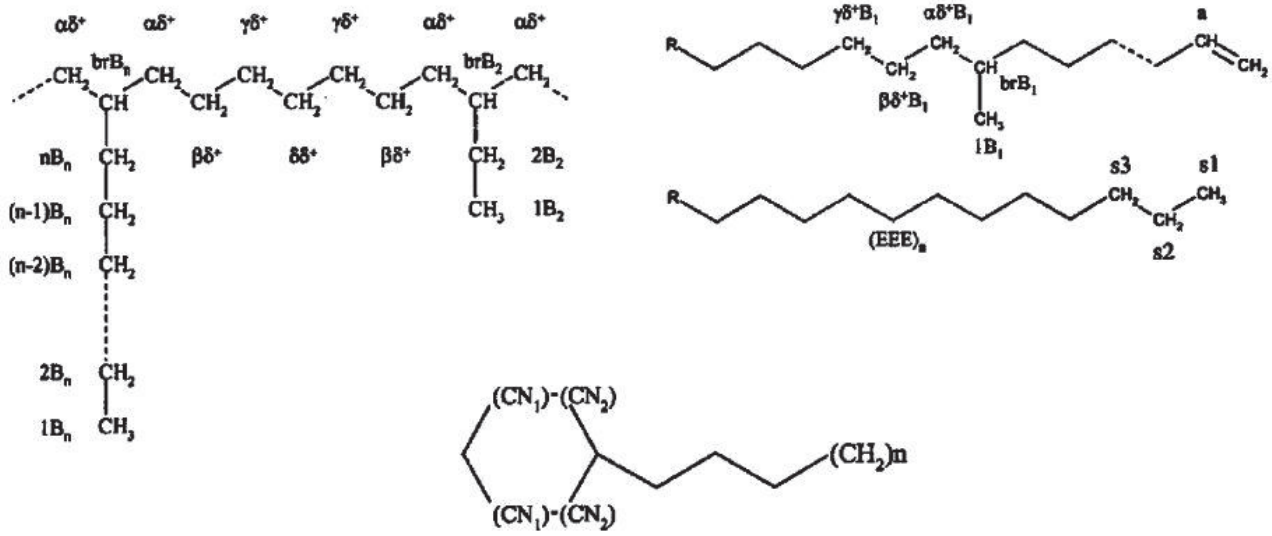


Figure A. 1 : Illustration des structures identifiables par RMN du ^{13}C [114]

Tableau A. 7 : Déplacement chimique correspondants aux structures de la Figure A. 1 [114]

Type	Label	Déplacement chimique (ppm)
Primaire (CH_3)	$1B_2$	11,5
	S_1	14,2
	$1B_{3\text{-}6}$	14,5
	$1B_1$	19,9
Secondaire (CH_2)	S_2	22,8
	CN_1	26,8
	$\beta\delta^+B_{3-6}$	27,21
	$\beta\delta^+B_1$	27,36
	$2B_2$	27,6
	EEEn	29,9
	$\gamma\delta^+B_1$	30,3
	$\gamma\delta^+B_{3-6}$	30,4
	S_3	32,0
	CN_2	33,8
	$\alpha\delta^+B_2$	34,0
	$\alpha\delta^+B_{3-6}$	34,5
	$\alpha\delta^+B_1$	37,5
Tertiaire (CH)	brB_1	33,1
	brB_2	34,8
	$brB_{3\text{-}6}$	37,0

Annexe F : Familles d'hydrocarbures identifiables par GC×GC

Tableau A. 8 : Liste non-exhaustive d'exemples de composés pour chacune des familles identifiables par GC×GC [35]

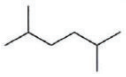
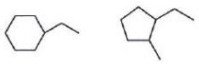
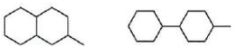
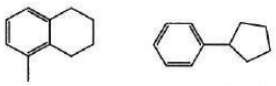
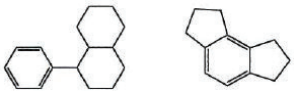
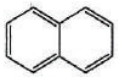
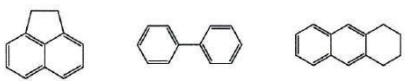
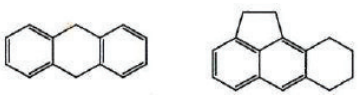
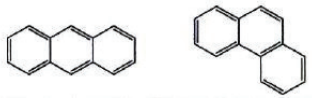
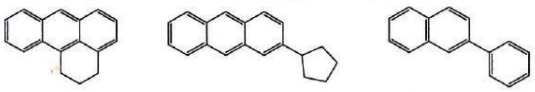
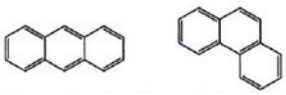
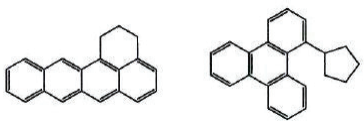
Formule brut	Famille chimique	Exemples de composés
n-C_nH_{2n+2}	N-paraffines	
i-C_nH_{2n+2}	Isoparaffines	
C_nH_{2n}	Mononaphtènes	
C_nH_{2n-2}	Dinaphtènes	
C_nH_{2n-4}	Trinaphtènes	
C_nH_{2n-6}	Monoaromatiques	
C_nH_{2n-8}	Naphténo-aromatiques	
C_nH_{2n-10}		
C_nH_{2n-12}	Diaromatiques	
C_nH_{2n-14}	Naphténo-diaromatiques	
C_nH_{2n-16}		

Tableau A. 9 : Liste non-exhaustive d'exemples de composés pour chacune des familles identifiables par GC×GC (suite) [35]

Formule brut	Famille chimique	Exemples de composés
C_nH_{2n-18}	Triaromatiques	
C_nH_{2n-20}		
C_nH_{2n-22}		
C_nH_{2n-24}	Tetraaromatiques	
C_nH_{2n-26}		

Annexe G : Lissage de données spectrales par méthode de Savitzky – Golay

La méthode de Savitzky-Golay permet de lisser ou de dériver une suite de résultats expérimentaux, d'intervalles réguliers, par une simple convolution avec une série de coefficients correspondant au degré du polynôme choisi et à l'opération souhaitée (simple lissage ou dérivation jusqu'à l'ordre 5) [148]. La méthode est explicitée ci-dessous :

Considérons une série de données régulièrement espacées $(f_i)_{i=-n,\dots,n}$ obtenues à partir d'une fonction f :

$$\boldsymbol{f_i = f(t_i)\ avec\ t_i = t_0 + i\Delta} \qquad \textbf{(Eq. A. 10)}$$

où Δ est une constante. La méthode de Savitzky-Golay consiste à remplacer chaque valeur f_i par la moyenne empirique des $(f_{i-n_i}, \dots, f_{i+n_i})$ notée g_i, *i.e.* :

$$\boldsymbol{g_i = \sum_{k=-n_i}^{n_i} \frac{1}{(2n_i+1)} f_{i+k}} \qquad \textbf{(Eq. A. 11)}$$

Où n_i est le nombre de point utilisé avant ou après la $i^{ème}$ valeur des données. Ce nombre est fixé par l'utilisateur au moment du lissage.

Annexe H : Méthodes de sélection de variables pour les modèles prédictifs

La théorie de l'analyse de sensibilité est en fait basée sur une approche stochastique appliquée à des ensembles déterministes. Considérons que l'on fixe un facteur x_i à la valeur m_i. On suppose alors que y peut être vue comme une variable aléatoire uniformément distribuée sur l'ensemble des valeurs que peut prendre la fonction $f(.,\dots,m_i,.,\dots)$ définie sur $[0;1]^{p-1}$. On note cette variable $(y|x_i = m_i)$. La quantité $\mathbb{E}(y|x_i = m_i)$ désigne alors la valeur moyenne de y lorsque x_i vaut m_i. La variable aléatoire $\mathbb{E}(y|x_i)$ définie sur $[0;1]$ par :

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x_i} =.) : \boldsymbol{m} \to \mathbb{E}(\boldsymbol{y}|\boldsymbol{x_i} = \boldsymbol{m}) \qquad \textbf{(Eq. A. 12)}$$

, décrit donc le comportement de la réponse y lorsqu'on fixe x_i à différentes valeurs. On peut ainsi l'assimilé à un estimateur de y qui dépend uniquement de x_i. On montre de plus que pour un descripteur x_i cet estimateur est optimal puisqu'il minimise l'erreur quadratique moyenne par rapport à y, *i.e.* [51] :

$$\mathbb{E}(\boldsymbol{y}|\boldsymbol{x_i}) = \min_{\boldsymbol{g}} \mathbb{E}(\boldsymbol{y} - \boldsymbol{g}(\boldsymbol{x_i}))^{\boldsymbol{2}} \qquad \textbf{(Eq. A. 13)}$$

et que le minimum atteint vaut [51]:

$$\mathbb{E}\big(\boldsymbol{y} - \mathbb{E}(\boldsymbol{y}|\boldsymbol{x_i})\big)^{\boldsymbol{2}} = \boldsymbol{Var}(\boldsymbol{y})\left(\boldsymbol{1} - \frac{\boldsymbol{Var}\big(\mathbb{E}(\boldsymbol{y}|\boldsymbol{x_i})\big)}{\boldsymbol{Var}(\boldsymbol{y})}\right) = \boldsymbol{Var}(\boldsymbol{y})(\boldsymbol{1} - \boldsymbol{S_i}) \qquad \textbf{(Eq. A. 14)}$$

L'erreur quadratique moyenne est donc une fonction décroissante

Ainsi, on dispose l'ensemble $\big(\mathbb{E}(y|x_i)\big)_{i=1,\dots,p}$ constitue une suite de p estimateurs de y.

Soit $i_0 \in \{1,\dots,p\}$ tel que x_{i_0} possède l'indice de sensibilité le plus élevé *i.e.* :

$$\boldsymbol{S_{i_0}} = \max_{\boldsymbol{i}\in\{\boldsymbol{1},\dots,\boldsymbol{p}\}} \boldsymbol{S_i} \qquad \textbf{(Eq. A. 15)}$$

$\mathbb{E}\big(y|x_{i_0}\big)$ est l'estimateur d'erreur quadratique moyenne minimale et celui qui décrit le mieux les variations de y. x_{i_0} correspond donc à la variable la plus influente.

Annexe I : PT en fonction de la teneur en n-paraffines de différentes espèces

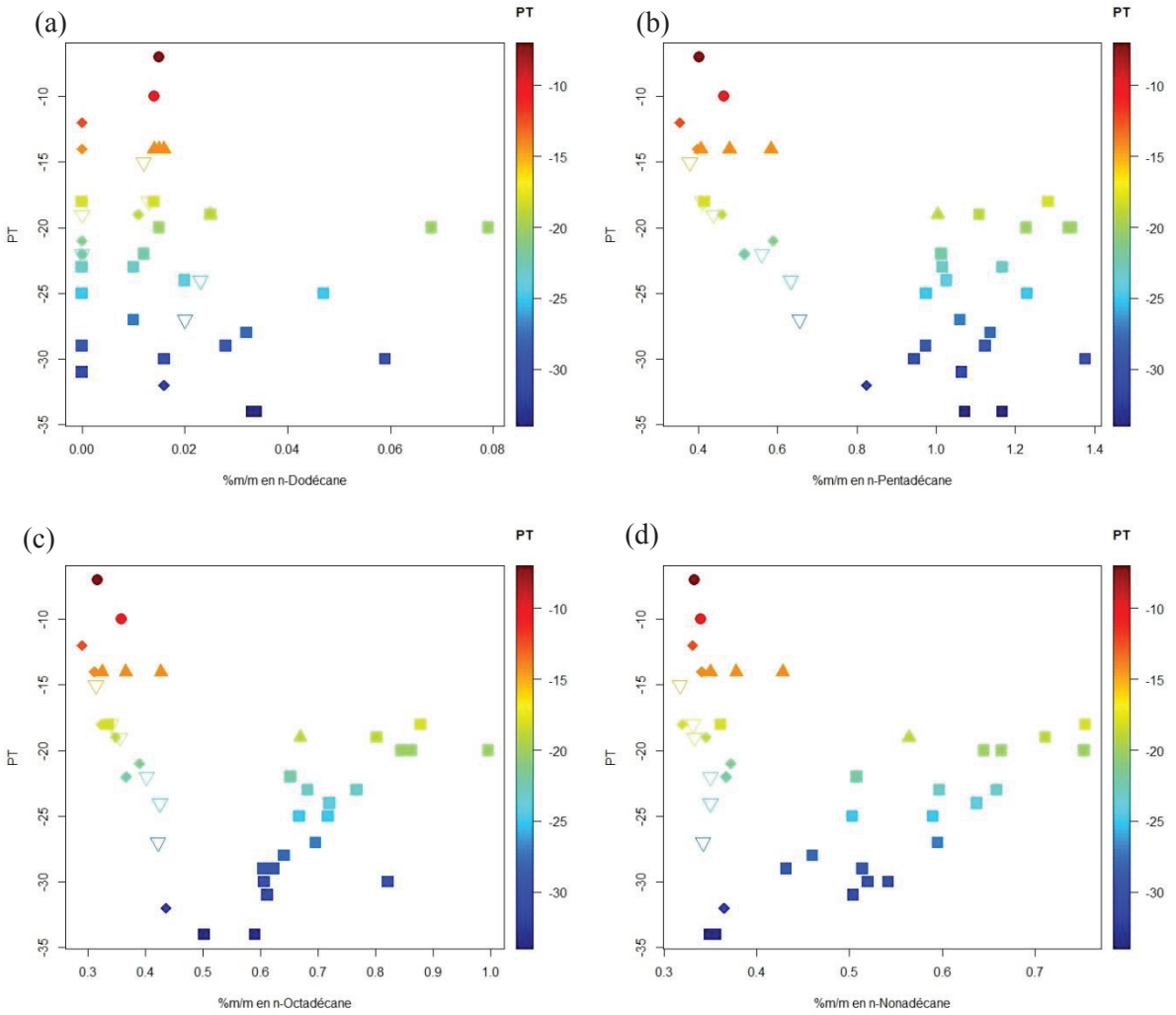


Figure A. 2 : Evolution du PT en fonction de la concentration de plusieurs n-paraffines contenues dans les gazoles ; a) en fonction de la concentration de n-Dodécane (n-C_{12}) ; b) en fonction de la concentration de n-Pentadécane (n-C_{15}) ; c) en fonction de la concentration de n-Octadécane (n-C_{18}) ; d) en fonction de la concentration de n-Nonadécane (n-C_{19})

Annexe J : Evolution du PT en fonction de la teneur en différentes familles d'hydrocarbures

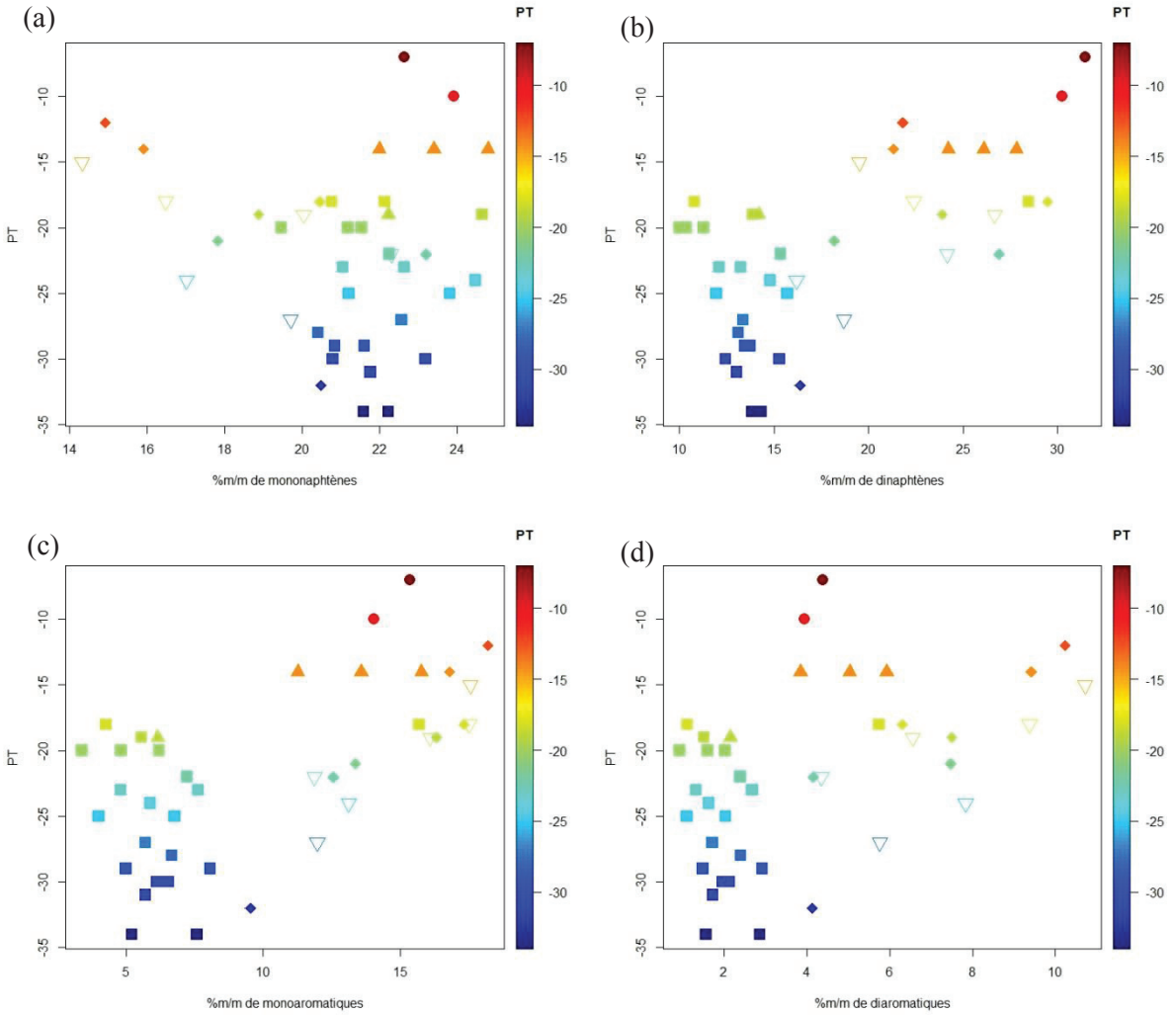


Figure A. 3 : Evolution du PT en fonction de la teneur de différentes familles d'hydrocarbures ; a) en fonction de la teneur en mononaphtènes ; b) en fonction de la teneur en dinaphtènes ; c) en fonction de la teneur en monoaromatiques ; d) en fonction de la teneur en diaromatique

Annexe K : Scores des échantillons sur les quatre premières composantes de l'ACP réalisée sur les données GC×GC des échantillons d'huile

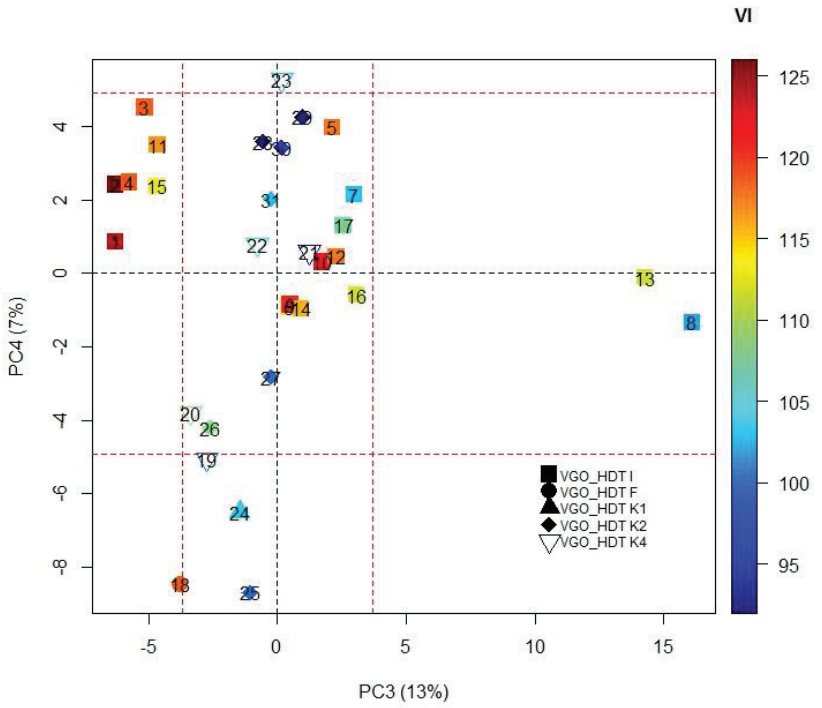


Figure A. 4 : Scores des échantillons d'huile caractérisés par GC×GC sur les composantes (PC3, PC4)

Annexe L : *Loadings* des variables chromatographiques sur les composantes principales de l'ACP réalisée sur les données GC×GC des échantillons d'huile

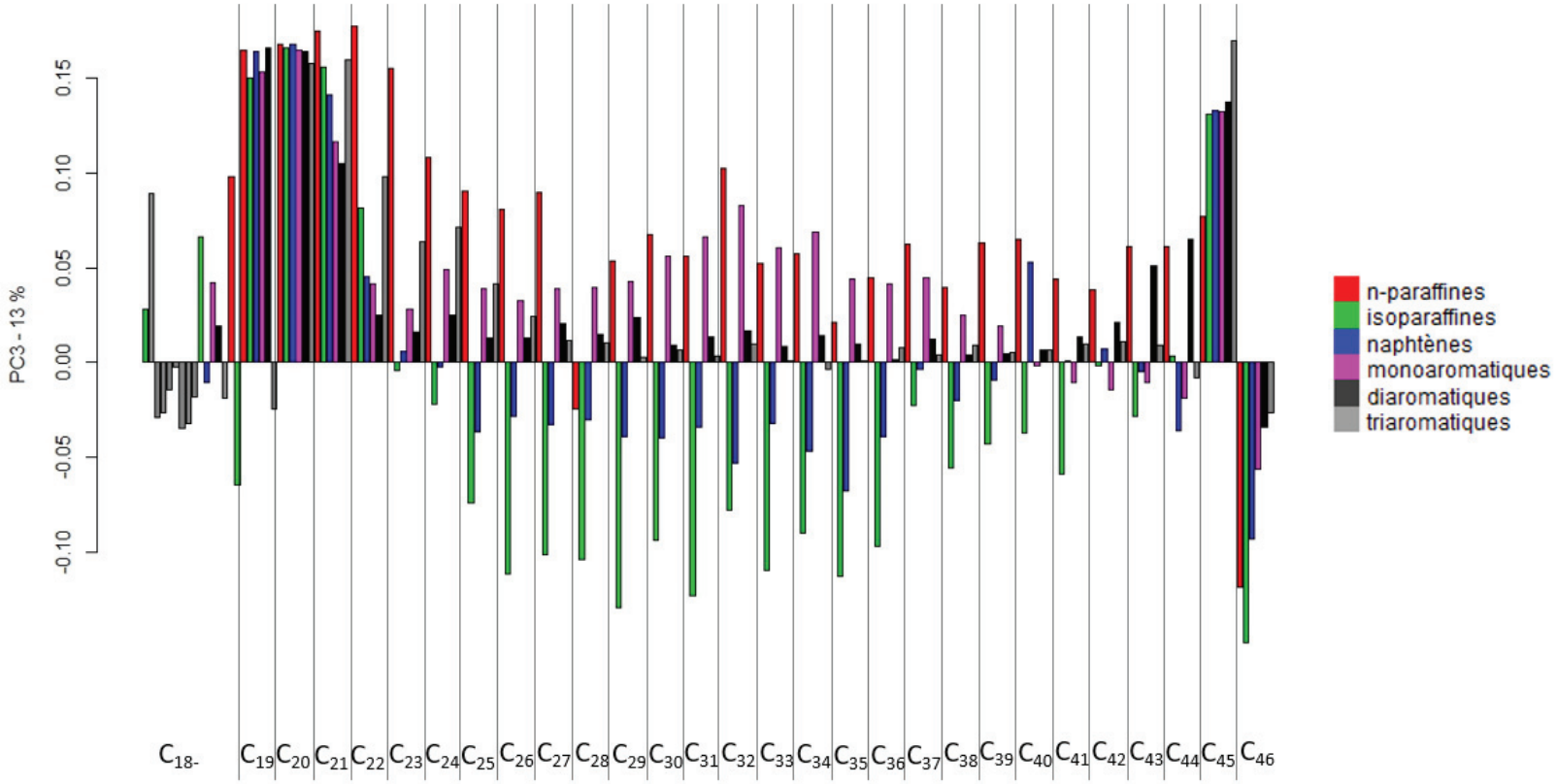


Figure A. 5 : *Loadings* des variables chromatographiques sur la composante PC3

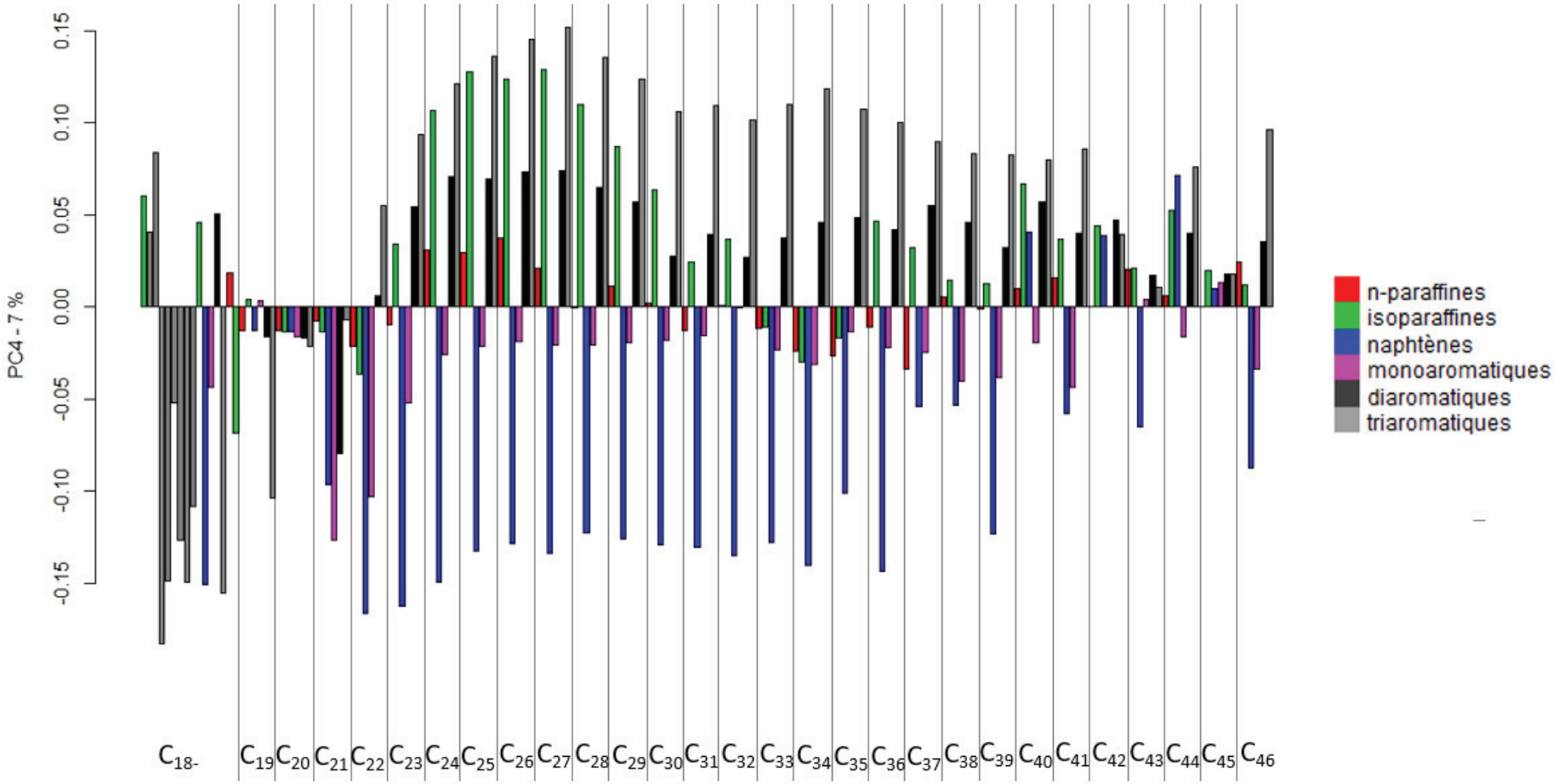


Figure A. 6 : *Loadings* des variables chromatographiques sur la composante PC4

Annexe M : *Review* sur les méthodes de détection d'*outliers* et application à la modélisation du VI de la coupe huile

Des travaux de *review* des méthodes de détection d'*outliers* ont été réalisés au cours de cette thèse. Différentes méthodes ont été recensées et tester sur un cas d'application.

A- Définitions et problèmes liés aux *outliers*

Barnett *et al*. [40] ont défini un *outlier* comme une observation qui apparait inconsistante avec le reste des données. La présence d'un *outlier* peut être la conséquence d'une contamination d'une autre distribution (bilan issu d'un procédé différent) ou d'une erreur de mesure ou de report de celle-ci. Les *outliers* ont un impact sur la qualité des modèles. Il est donc indispensable de les détecter pour les retirer de la base de données. Les effets des *outliers* se présentent généralement sous deux formes : l'effet masquant qui caractérise la non détection d'un ou plusieurs *outliers* en raison de la présence d'autres ; l'effet de remplissage qui traduit le fait qu'une observation correcte soit vue comme *outlier*. Ces deux phénomènes sont généralement dus à des effets de compensation ou à un déplacement significatif du centre de gravité du nuage de point.

Face à ce phénomène, deux stratégies peuvent être adoptées (Figure A. 7) :

1. La détection des *outliers* par différentes méthodes en vue d'une exclusion
2. L'utilisation de méthodes de modélisation dites robustes qui permettent de réduire l'influence d'un point isolé.

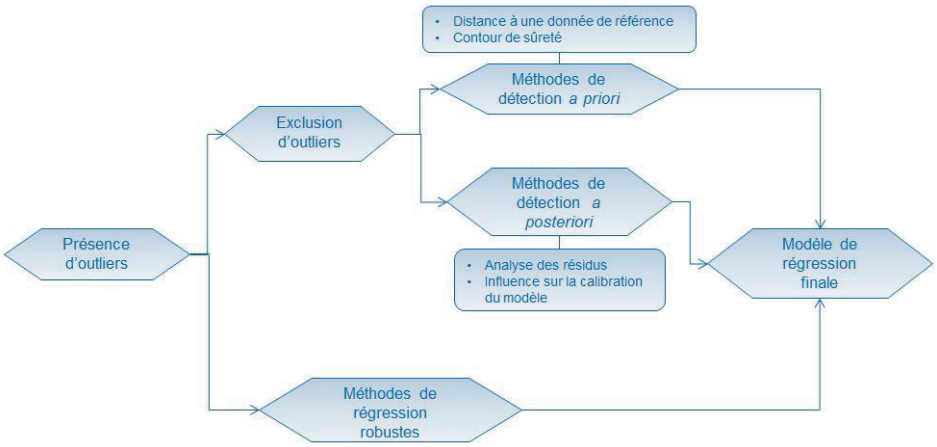


Figure A. 7 : Schéma récapitulatif des stratégies face aux *outliers*

Dans ce qui suit, nous nous limiterons essentiellement aux méthodes de détection d'*outlier*.

B- Méthodes de détection d'*outliers*

Une étude bibliographique des méthodes de détection d'*outliers* dans les bases de données a été réalisée. Elle a abouti au recensement de plusieurs méthodes. Ces méthodes ont été classées suivant

deux catégories : les méthodes de détection non supervisées ou méthodes *a priori* qui reposent sur une analyse géométrique des données ; les méthodes supervisées ou diagnostic *a posteriori* qui nécessitent de prédéfinir un modèle. Ces méthodes sont récapitulées dans les

B.1- Méthodes de détection non supervisées

La plupart des méthodes non supervisées sont basées sur le concept de la **distance de Mahalanobis** qui évalue la distance d'un point de la base de données au centre de gravité du nuage global [174]. Sous l'hypothèse que les variables caractéristiques sont toutes normalement distribuées, la distance de Mahalanobis doit être inférieure au quantile d'ordre 0,95 d'une loi du χ^2 à p degrés de liberté (p étant la dimension de l'espace d'étude) [174]. La méthode de Mahalanobis classique qui utilise les estimateurs empiriques de la moyenne et la covariance étant sensible à la présence de multiples *outliers* (déplacement significatif du centre de gravité), plusieurs auteurs ont cherché à y remédier en introduisant des **estimateurs robustes**. On appelle notamment estimateur robuste tout estimateur qui ne dépend que très faiblement d'une donnée isolée. Les estimateurs MCD (Minimum Covariance Determinant) , MVE (Minimum Volume Ellipsoid), RHM (Resampling Half Mean) et SHV (Smallest Half Volume) ont ainsi été développés [175].

D'autres méthodes non supervisées ont également été recensées : la méthode des fonctions de potentiel [176], la méthode de Dixon qui utilise des tables statistiques. Cette dernière n'est cependant pas pertinente dans le cas de modèles multidimensionnels ou multivariés car elle requiert d'effectuer un trop grand nombre de tests. L'idée des fonctions de potentiel est de fournir une mesure de l'affluence au tour d'un point donné. Pour cela, on définit le potentiel élémentaire ϕ induit par un point x_i en un point x_j comme une fonction de la distance entre ces deux points. On en déduit alors le potentiel global en x_j comme la moyenne des potentiels élémentaires induits par les points de la base de données en ce point. Il existe plusieurs formes de potentiels élémentaires parmi lesquelles [176] :

- le potentiel gaussien défini par :

$$\boldsymbol{\phi(x_i, x_j) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2s^2}\left(x_i - x_j\right)^2\right)} \qquad \textbf{(Eq. A. 16)}$$

- le potentiel triangulaire :

$$\boldsymbol{\phi(x_i, x_j) = \begin{cases} 1 - \left|\frac{x_i - x_j}{s}\right| \text{ si } \left|\frac{x_i - x_j}{s}\right| \leq 1 \\ 0 \text{ sinon} \end{cases}} \qquad \textbf{(Eq. A. 17)}$$

Un point sera considéré *outlier* si son potentiel global est relativement faible (ou nul) par rapport à une certaine valeur seuil. Les différentes méthodes non supervisées sont reportées dans le Tableau A. 10.

Tableau A. 10 : Récapitulatif des méthodes de détection d'*outliers* non supervisées

Méthodes	Statistique	Valeur critique	Signification mathématique
Méthode de Mahalanobis classique	$d_i = (x_i - \bar{x})\Sigma^{-1}(x_i - \bar{x})'$ Σ ↔ matrice de covariance des variables $\bar{x}$ ↔ vecteur moyen	χ_p^2	Distance de l'observation par rapport au centre de gravité du nuage formé par l'ensemble des points
Méthode Minimum Covariance Determinant (MCD)	$d_i = (x_i - \overline{x_{MCD}})\Sigma_{MCD}{}^{-1}(x_i - \overline{x_{MCD}})'$ Σ_{MCD} ↔ matrice de covariance des variables sur sous échantillon MCD $\overline{x_{MCD}}$ ↔ vecteur moyen sur sous échantillon MCD	χ_p^2	Variation relative du volume l'ellipsoïde de confiance des paramètres suivant que l'observation i est ou non prise en compte
Méthode Minimum Volume Ellipsoid (MVE)	$d_i = (x_i - \overline{x_{MVE}})\Sigma_{MVE}{}^{-1}(x_i - \overline{x_{MVE}})'$ Σ_{MVE} ↔ matrice de covariance des variables sur sous échantillon MVE $\overline{x_{MVE}}$ ↔ vecteur moyen sur sous échantillon MVE	χ_p^2	Vérification de l'hypothèse de normalité des résidus
Méthode Resampling Half Means (RHM)	$d_i = (x_i - \overline{x_{RHM}})\Sigma_{RHM}{}^{-1}(x_i - \overline{x_{RHM}})'$ Σ_{RHM} ↔ matrice de covariance des variables sur sous échantillon RHM $\overline{x_{RHM}}$ ↔ vecteur moyen sur sous échantillon RHM	χ_p^2	Vérification de l'hypothèse de normalité des résidus
Méthode Smallest Half Volume (SHV)	$d_i = (x_i - \overline{x_{SHV}})\Sigma_{SHV}{}^{-1}(x_i - \overline{x_{SHV}})'$ Σ_{SHV} ↔ matrice de covariance des variables sur sous échantillon SHV $\overline{x_{SHV}}$ ↔ vecteur moyen sur sous échantillon SHV	χ_p^2	Variation relative du volume l'ellipsoïde de confiance des paramètres suivant que l'observation i est ou non prise en compte
Fonction de potentiel	$f(x_j) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i, x_j)$ $\phi(x_i, x_j)$ ↔ potentiel induit par le point i au point j	Dépend du choix de ϕ	Le potentiel en un point qui mesure la densité locale en ce point. Plus un point est isolé plus le potentiel induit est faible
Q test de Dixon	Voir tables statistiques de Dixon	Voir tables statistiques	Différence relative entre deux observations supposées appartenir à une même distribution

χ_p^2 : quantile d'une loi du khi-deux à p degrés de liberté

B.2- Méthodes de détection supervisées

Les méthodes de détection dites supervisées (Tableau A. 11) reposent sur des analyses statistiques de résidus obtenus à partir d'un modèle prédéfini (résidus standardisés, résidus de Student, *etc.*) et sur la notion d'influence des observations sur la phase d'apprentissage du modèle (Effet de levier). Ces méthodes sont en général définies pour des modèles de régression linéaire multiple mais peuvent être généralisées à des cas de régression non linéaire. La plupart des méthodes présentées ci-

dessus ont été comparées sur un jeu de données simulées par Walczak et Massart [177]. Ils ont noté que **les méthodes les plus performantes à savoir le cov ratio (CVR), la statistique de Andrews-Pregibon et (p*) et le critère de Welsh-Kuh (WK) donnent des résultats très proches et sont relativement sensibles à l'effet de remplissage**.

Tableau A. 11 : Récapitulatif des méthodes de détection d'*outliers* supervisées

Méthodes	Statistique	Valeur critique	Signification mathématique
Effet de Levier (p)	$H_{i,i} = h_{ii} = p_i$	$\frac{2p}{n}$	$\frac{\partial \widehat{y_i}}{\partial y_i}$
Statistique de Andrews-Pregibon (p*)	$p_i^* = 1 - AP_i$ $AP_i = 1 - p_i - \frac{\hat{\varepsilon}_i^2}{\hat{\varepsilon}^T\hat{\varepsilon}}$	$\frac{2(p+1)}{n}$	Variation relative du volume de l'ellipsoïde de confiance des paramètres suivant que l'observation i est ou non prise en compte
Résidus standardisés (t)	$t_i = \frac{\widehat{\varepsilon_i}}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \widehat{\varepsilon_i}^2$	Test d'appartenance à une loi normale centrée réduite	Vérification de l'hypothèse de normalité des résidus
Résidus « studentisés » (t*)	$t_i^* = t_i\sqrt{(n-p-1)/(n-p-t_i^2)}$	$t_{n-p-1,1-\alpha/2}$	Vérification de l'hypothèse de normalité des résidus
Cov Ratio (CVR)	$CVR_i = \left[\frac{n-p-t_i^2}{n-p-1}\right]^p /(1-p_i)$	$\lvert CVR_i - 1\rvert > \frac{3p}{n}$	Variation relative du volume l'ellipsoïde de confiance des paramètres suivant que l'observation i est ou non prise en compte
Distance de Cook (C)	$C_i = \frac{(\hat{\beta}-\widehat{\beta_{(i)}})X'X(\hat{\beta}-\widehat{\beta_{(i)}})}{\hat{\sigma}^2 p} = \frac{h_{ii}}{p(1-h_{ii})}t_i^2$[41]	$f_{p,n-p,1-\alpha}$	Test d'appartenance du vecteur des paramètres obtenus sans prise en compte de l'observation i
Distance de Welsh-Kuh (WK)	$WK_i = \lvert t_i^*\rvert\sqrt{h_{ii}(1-h_{ii})}$	$2\sqrt{(\frac{p}{n})}$	Couplage résidus studentisés et effet de levier
Distance de Cook-Welsh-Kuh (C*)	$C_i^* = WK_i\sqrt{(n-p)/p}$	$2\sqrt{(n-p)/n}$	Couplage critère de Welsh-Kuh et critère de Cook

A la suite des observations recueillies au cours de notre étude bibliographique, nous avons choisi de comparer cinq méthodes de détection d'*outliers* :

- La distance de Mahalanobis classique,

- La distance de Mahalanobis robuste avec estimateur MCD,
- la méthode des fonctions de potentiel,
- le cov ratio
- la statistique de Andrews-Pregibon.

C- Application des méthodes de détection d'*outliers* à la modélisation du VI des huiles de base

Les méthodes sélectionnées dans le paragraphe précédent ont été testées sur un jeu de données simulées. Dans la suite, nous détaillons les traitements qui ont été effectués. Ces derniers sont récapitulés sur le schéma de la Figure A. 8 :

1. Préparation de la base de données
2. Génération aléatoire d'*outliers*
3. Application des méthodes de détection sélectionnées
4. Comparaison des performances

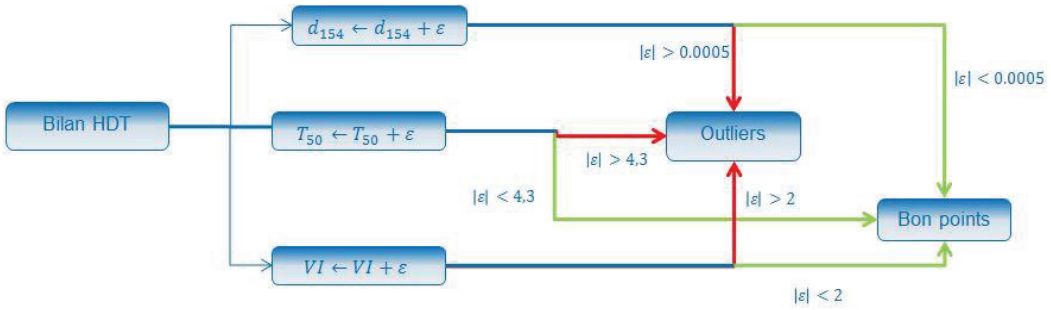


Figure A. 8 : Stratégie de bruitage des données analytiques pour la prédiction du VI

C.1- Préparation de la base de données

Une base de données de 46 bilans HDT a été constituée. Ces bilans proviennent essentiellement de deux charges DSV : la charge VGO K et la charge VGO L qui ont des caractéristiques très proches. Ce choix permet notamment de réduire le nombre de descripteurs, le modèle de prédiction du VI ne nécessitant alors que les variables T_{50} et d_{15}^4.

A partir de cette base de données (considérée comme propre), une base de données bruitées a été générée **de sorte qu'elle contienne environ 5% d'*outliers***. Pour cela, chaque point de la base de données initiale a été dupliqué un certain nombre de fois en rajoutant un **bruit blanc gaussien d'un écart-type équivalent à** $\frac{R}{\sqrt{2}}$ où R est la reproductibilité de la mesure de référence.

Dans l'étude présentée ici, une base de données de 920 bilans contenant 47 *outliers* de mesure a été générée. Les données sont représentées sur la Figure A. 9 dans l'espace (d_{154},T_{50},VI). L'axe vertical fait référence au VI. Les variables ont été normalisées. Les points corrects sont en vert et les

outliers en rouge. Les cinq méthodes de détection d'*outliers* sélectionnées ont chacune été appliquées sur ce jeu de données.

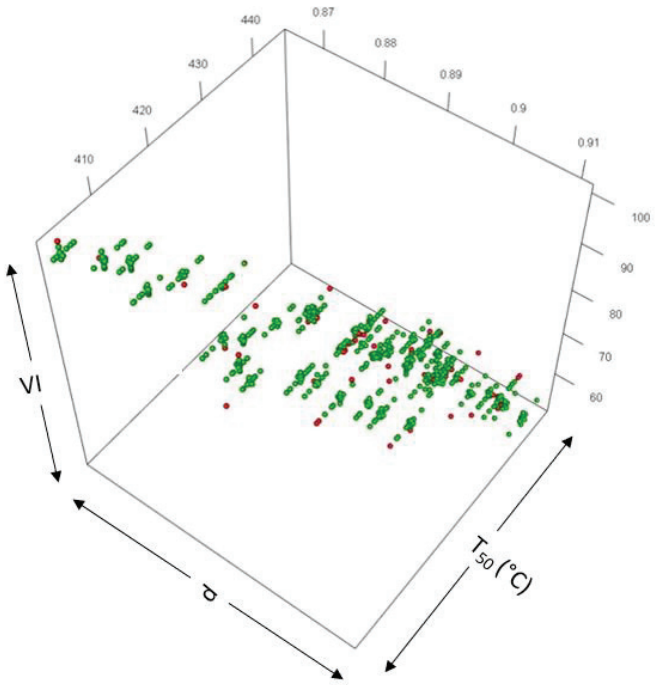


Figure A. 9 : Base de données bruitée contenant 920 observations ; en vert → points corrects ; en rouge → *outliers*

C.2 – Discussions et résultats

C.2.1 – Application du critère de Mahalanobis

Les critères de Mahalanobis classique et robuste (MCD) ont été appliqués pour détecter les *outliers* de la base de données décrites au paragraphe précédent. Les ellipsoïdes de confiance obtenues dans chacun des cas sont représentées sur la Figure A. 10. On note que le critère de Mahalanobis classique (Figure A. 10a) ne détecte que deux *outliers* (points rouges à l'extérieur de l'ellipsoïde) sur 47. On observe également un effet de remplissage puisque certains points verts sont à l'extérieur de l'ellipsoïde de confiance et sont donc vus comme *outliers*. Concernant le critère robuste MCD (Figure A. 10b), 3 *outliers* sont détectés. Par contre, ce critère apparait beaucoup plus sévère puisqu'on observe l'exclusion d'une bonne quantité de points verts.

De manière générale, la forme ellipsoïdale de la zone de confiance imposée par le critère de Mahalanobis (classique et robuste) est peu adaptée à une base de données complexe. Cette méthode manque de flexibilité.

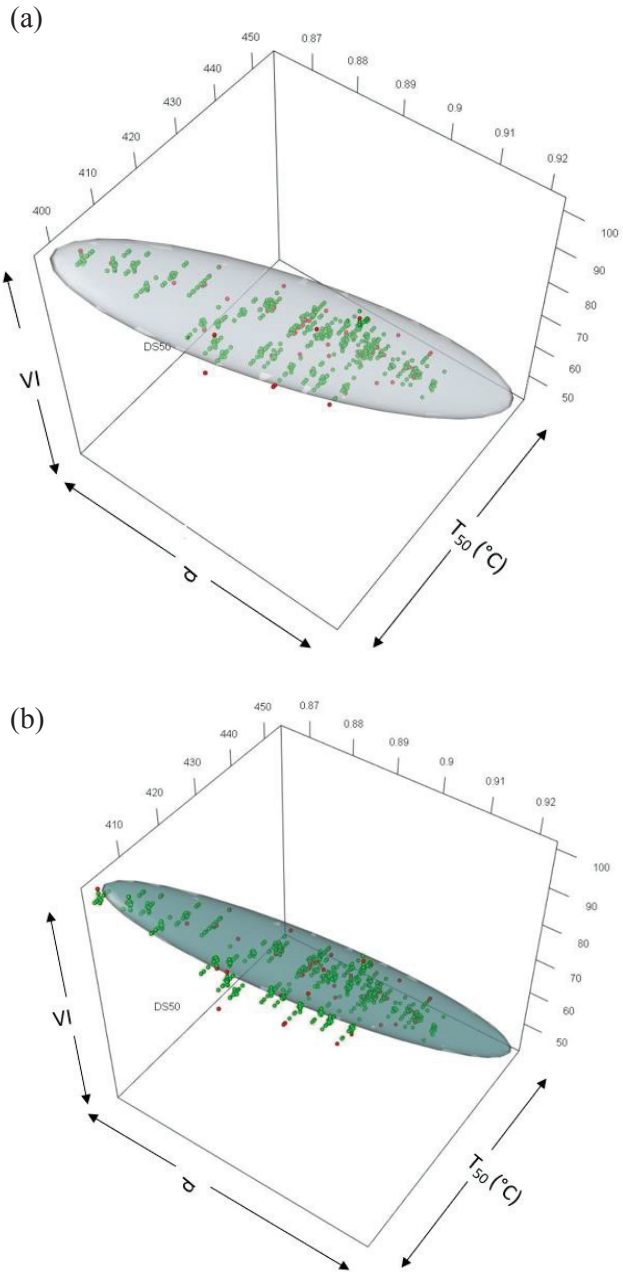


Figure A. 10 : Ellipsoïdes de confiance obtenues par critère de Mahalanobis; a) – cas classique ; b) – cas robuste MCD

C.2.2- Application des fonctions de potentiel

La fonction de potentiel choisie pour cette étude est la fonction triangulaire. Pour détecter les *outliers* dans la base de données, un algorithme *leave-one-out* consistant à calculer le potentiel induit en chaque point par le reste de la base de données a été appliqué (un point est considéré comme *outlier* par la méthode si son potentiel induit est nul). Les résultats sont représentés sur la Figure 85. Seulement 5 *outliers* sont détectés et là encore on observe clairement des effets de remplissage. Le *leave-one-out* n'est donc pas suffisamment efficace.

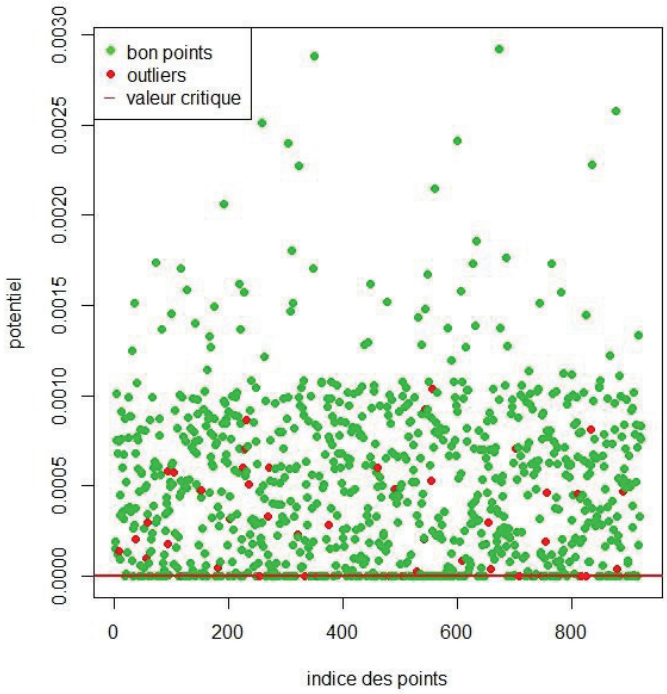


Figure 85 : Potentiel induit en chaque point par le reste de la base de données.

La méthode des fonctions de potentiel est plus sélective et plus flexible que le critère de Mahalanobis. En effet, elle permet de définir des zones de confiance qui peuvent s'adapter à des nuages de points plus complexe (présence de clusters par exemple) du fait du caractère local de la méthode. Le contour délimité par la fonction de potentiel dans le cas présent est illustré sur la Figure 86. Cela montre que les fonctions de potentiel peuvent être utilisées pour définir des domaines de prédictibilité. La définition d'une valeur seuil reste cependant un problème ouvert.

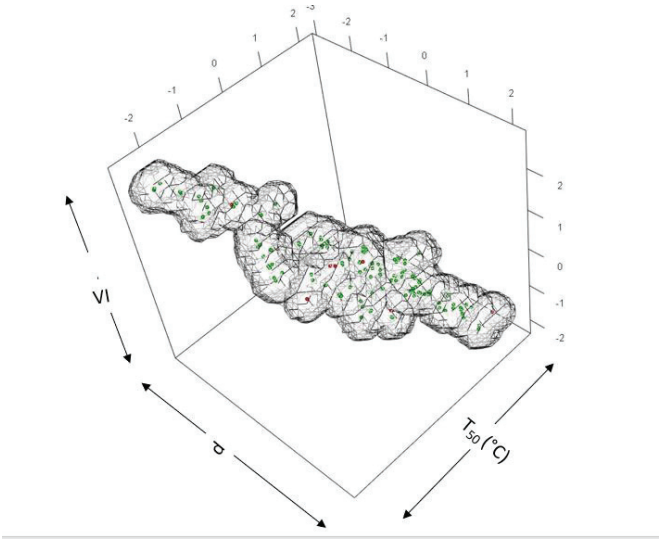


Figure 86 : Contour de sûreté délimité par l'utilisation de fonctions de potentiel

C.2.3 – Application de méthodes diagnostic *a posteriori*

Pour l'étude *a posteriori* seul les critères du cov ratio et de l'effet de levier pénalisé ont été retenus. Ces méthodes requièrent de prédéfinir un modèle. Dans cette étude, le modèle choisi est linéaire bivarié. La surface de réponse obtenue à partir de la base de données bruitées est représentée sur la Figure A. 11 (plan transverse incliné). La couleur attribuée à chaque point fait référence à la mesure du cov ratio. Les points bleus correspondent à une influence normale, tandis que les points rouges correspondent à une influence forte. On note que les points de plus forte influence (au sens du cov ratio) sont la plupart du temps ceux qui sont les plus écartés de la surface de réponse ou du centre de gravité du nuage de points, ce qui est logique dans le cas d'un modèle linéaire. Cependant il y a des zones de forte influence moins évidentes. C'est le cas de la zone entourée en vert, où l'on distingue un point rouge (donc de forte influence) au milieu de point d'influence normale. Cela montre que la notion d'influence peut être très spécifique. Les mêmes observations peuvent être faites sur le critère de l'effet de levier.

La Figure A. 12a montre les mesures d'influence associées à chaque point de la base de données selon le critère de l'effet de levier. La Figure A. 12b présente les mesures d'influence obtenues dans le cas du cov ratio. Les points situés au-dessus de ligne rouge (qui représente la valeur critique) sont vus comme *outliers* par la méthode considérée. On note que :

- L'effet de levier est moins performant que le cov ratio pour la détection d'*outliers* ;
- Le critère du cov ratio est particulièrement pessimiste ; on observe un fort effet de remplissage, c'est-à-dire que les points corrects sont détectés comme *outliers* ;

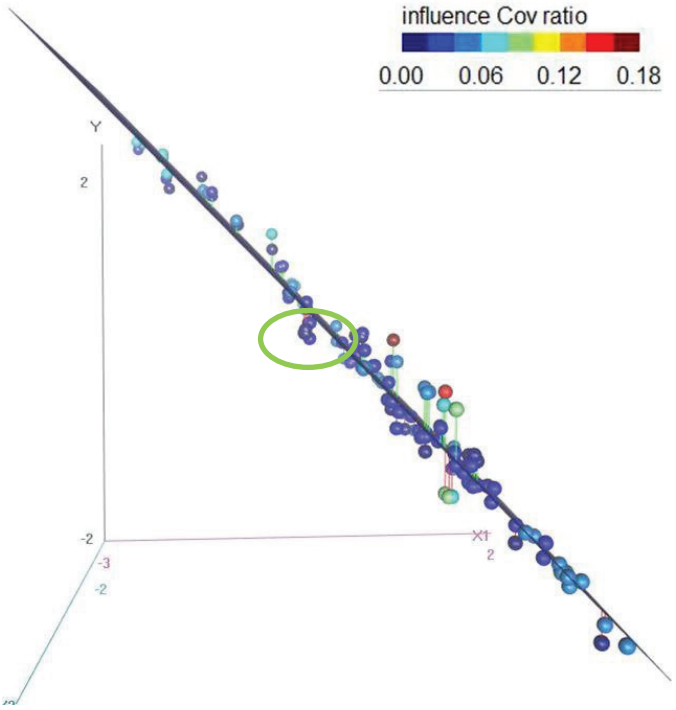
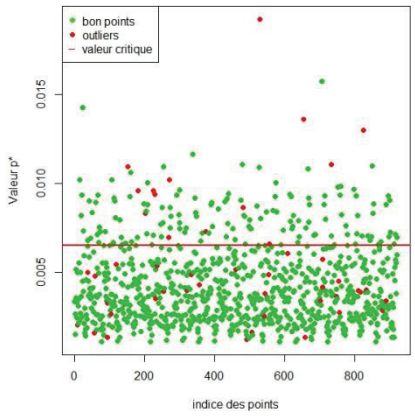


Figure A. 11 : Mesure d'influence des points de la base de données par méthode du cov ratio

(a)



(b)

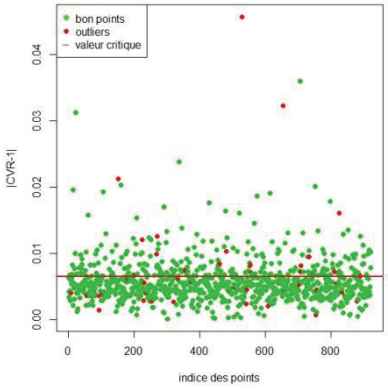


Figure A. 12 : Mesures d'influence des points de la base de données a) par effet de levier ; b) par cov ratio

C.2.4 – Comparaisons des méthodes

Le Tableau A. 12 ci-dessous confirme les observations faites précédemment. Le critère du cov ratio, l'effet de levier pénalisé et la méthode des fonctions de potentiel permettent de détecter respectivement 20,14 et 12 sur 47 *outliers*. Cependant on observe que le cov ratio est très pessimiste puisqu'il détecte 252 points douteux pour les fonctions de potentiel. Le critère de Mahalanobis classique est le moins performant pour ce qui est de la détection d'*outlier*. Il est aussi le moins sensible à l'effet de remplissage. **Il y a donc un compromis à trouver entre effet masquant et effet de remplissage**.

Tableau A. 12 : Résultats de l'application des méthodes de détection d'*outliers* à la base de données bruitées ex-HDT

Critère	Bons points classifiés comme tels	Bons points vus comme *outliers*	*Outliers* classifiés comme tels	*Outliers* vus comme bons points
Mahalanobis classique	856/873	17/873	6/47	41/47
Mahalanobis robuste	719/873	154/873	12/47	35/47
Fonction de potentiel	846/873	27/873	12/47	35/47
Cov ratio	621/873	252/873	20/47	27/47
Effet de levier pénalisé	853/873	20/873	14/47	34/47

C.3 – Conclusion

En résumé, cinq méthodes de détection d'*outliers* ont été appliquées à la prédiction du VI de la coupe huile en sortie HDT. Les résultats obtenus montrent que :

- les méthodes de Mahalanobis sont peu flexibles et ne permettent pas de détecter suffisamment d'*outliers* ;
- les fonctions de potentiel offrent un critère plus flexible et plus adapté à des bases de données complexes ;
- le cov ratio permet de détecter le plus grand nombre d'*outlier*, mais semble assez pessimiste ;
- l'effet de Levier pénalisé détecte moins d'*outliers* que le cov ratio mais est moins sensible à l'effet de remplissage.

De manière générale, la difficulté pour les méthodes qui permettent de détecter le plus grand nombre de vrais *outliers* (cov ratio, effet de levier pénalisé) est de trouver le bon compromis entre effet masquant et effet de remplissage. Les valeurs critiques théoriques associées à ces critères ne sont peut-être tout simplement pas généralisables à des jeux de données divers.

Pour les méthodes de détection supervisées c'est surtout la notion d'influence qui est remise en cause. **Dans le cas du cov ratio l'introduction d'un bruit analytique ne change pas significativement la mesure d'influence des points de la base de calibration**. Ce qui explique le nombre important de points exclus. **Un meilleur compromis est respecté dans le cas de l'effet de levier pénalisé**.

Pour les méthodes *a priori*, la distance de Mahalanobis classique n'est pas suffisamment sensible dans le cas des propriétés produits. **L'estimateur robuste (MCD) permet de réduire taille de la zone de sûreté sans tenir compte de la géométrie du nuage de points**, ce qui explique le grand nombre

de faux *outliers* détectés. **Cet aspect est mieux pris en compte dans le cas des fonctions de potentiel dont le caractère local offre plus de flexibilité**. L'utilisation d'un potentiel gaussien et le réajustement de la valeur seuil peut permettre d'améliorer le compromis entre effet masquant et effet de remplissage.