UNIVERSITE DE LILLE 1 – SCIENCES ET TECHNOLOGIES

ECOLE DOCTORALE DE SCIENCES DE LA MATIERE, DU RAYONNEMENT ET DE L'ENVIRONNEMENT

THESE

EN VUE D'OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITE DE LILLE 1

DISCIPLINE :

OPTIQUE, LASERS, PHYSICO-CHIMIE ET ATMOSPHERE

Présentée et soutenue publiquement par

Jérémy LAXALDE

le 16 Janvier 2012

ANALYSE DES PRODUITS LOURDS DU PETROLE PAR SPECTROSCOPIE INFRAROUGE

Rapporteurs :

Dr. Véronique BELLON-MAUREL	CEMAGREF, Montpellier
Pr. Philippe GIAMARCHI	Université de Bretagne Occidentale, Brest

Examinateurs

Pr. Douglas RUTLEDGE	AgroParistech, Paris
Dr. Sophãnara DE LOPEZ	TOTAL, Puteaux
Dr. François WAHL	IFP Energies Nouvelles, Solaize
Dr. Noémie CAILLOL	IFP Energies Nouvelles, Solaize

Directeur de thèse

Pr. Ludovic DUPONCHEL	Université Lille 1, Villeneuve d'Ascq

Co-directeur de thèse

Dr. Cyril RUCKEBUSCH	Université Lille 1, Villeneuve d'Ascq

Remerciements

Ce travail de thèse est le fruit d'une convention CIFRE nourrie par une collaboration entre l'université de Lille 1 et IFP Energies Nouvelles. Je tiens donc, tout d'abord, à remercier M. Guy Buntinx, directeur du LASIR, ainsi que M. Thierry Becue et Mme Nathalie Schildknecht, respectivement responsables à IFP Energies Nouvelles de la direction "Physique et analyse" et du département "Caractérisation des produits", d'avoir permis la réalisation de cette étude au sein de leurs établissements.

Je suis également honoré que les professeurs Véronique Bellon-Maurel et Philippe Giamarchi aient accepté d'être rapporteurs de cette thèse. Je les remercie pour leur lecture attentive de ce manuscrit, leurs remarques et pour l'intérêt qu'ils ont porté à ce travail. Ma gratitude va également au Professeur Douglas Rutledge et au Dr. Sophãnara de Lopez qui m'ont fait le plaisir d'accepter d'être, respectivement, président de mon jury et examinateur de ce travail de thèse.

Je tiens maintenant à exprimer ma sympathie et ma profonde reconnaissance à mes encadrants, qui au delà de m'avoir permis de mener à bien ce travail de thèse, m'ont initié au monde de la recherche et de l'entreprise et, ainsi de progresser personnellement et professionnellement. Je remercie tout d'abord Ludovic Duponchel, Professeur de l'université de Lille 1, d'avoir accepté de diriger ces travaux, pour son investissement et pour les précieux conseils qu'il a su me prodiguer durant ces trois années de thèse. Je tiens également à exprimer ma gratitude à Cyril Ruckebusch, Maître de conférence HDR à l'université de Lille 1, pour ses compétences multiples, pour sa grande contribution

dans la mise en forme et la valorisation des résultats obtenus lors de ces travaux. Je le remercie également pour la rigueur et la patience dont il a fait preuve durant les périodes rédactionnelles. Je tiens à remercier Noémie Caillol, Chargée de recherche à IFP Energies Nouvelles, tout d'abord pour sa part déterminante dans ma prise de décision de m'investir dans ce projet. De plus, je tiens également à la remercier pour sa disponibilité scientifique et humaine de tous les instants ainsi que pour sa volonté inébranlable tout au long de ces trois années. Enfin, je voudrais saluer François Wahl, Ingénieur à IFP Energies Nouvelles, d'avoir participé à mon encadrement de manière efficace et bienveillante et pour ses conseils formateurs en écriture scientifique.

Je veux également remercier toutes les personnes du département "Caractérisation des produits" qui m'ont aidé lors de mes travaux expérimentaux. Un grand merci à Pierre Paul et les membres du laboratoire "Analyses pétrolières" pour leur contribution lors de la collecte des échantillons de produits lourds. Je remercie également Olivier Delpoux, David Gonçalves, Corinne Sagnard et Sophie Bailly ainsi qu' Anne-Agathe Quoineaud, Laurent Lemaitre, Emmanuel Soyer, Mathieu Vidalie et Emmanuelle Sorbier-Richard de m'être venu amicalement en aide lors du développement des protocoles d'acquisition des spectres MIR et PIR. Un grand merci également à Olivier Devos pour son importante contribution lors de travaux d'optimisation des modèles chimiométriques par algorithmes génétiques.

Je veux maintenant adresser mes remerciements à toutes les personnes que j'ai pu cotoyer au cours de ces trois années de thèse. Tout d'abord, Je tiens à adresser ma reconnaissance aux membres de la colocation de Ronchin, Rémi, Nico, Ben et Alice, pour leur accueil dans le Nord et sans qui, mes déplacements à Lille aurait été plus compliqués mais surtout beaucoup moins conviviaux ! De plus, je remercie les personnes qui ont dû me supporter au jour le jour, sur une durée plus ou moins longue, dans le bureau 12DOL/1F19 : Fabien, Badaoui, Laure, Thomas et Vincent. Les moments passés dans ce bureau resteront d'excellents souvenirs ! Peu adepte du café lors de mon arrivée, j'en ai développé une forte addiction ! Cela, je le dois en plus grande partie à Matthieu et Florent, deux Monsieur de la machine à café, à qui je veux tout particulièrement adresser mes amitiés. Bien entendu, ils n'étaient pas les seuls. Je tiens donc à saluer également

tous ceux qui ont gravité, tout comme moi, autour de ce bijou de technologie, où l'on trouve réconfort et motivation dans les moments difficiles : Luc (pour les pains au choc), Mimmo, Olivier, Philippe, El Mifdol, Flo, Laurent, Laurent le Belge, Johan, Anthony, Dimitrius, Viet-Dung, Paolo, Daniel, Eric et j'en oublies!!!

Des travaux de thèse sont une mobilisation de tous les instants. Les moments de répit ne doivent donc pas être gachés et donc être partagés avec des personnes de qualité. C'est pourquoi, je tiens à saluer tout d'abord tous les bretons qui ont résidé à Lyon et avec qui j'ai passé d'agréables moments : Teuf, Mat, Claire, Pica, Aurélien, Sophie, Dam et Coco. Pour compléter ces remerciements je me dois de citer mes amis. Un grand merci à Glenn, Jérém, Kev, Elodie, Cédric, Marie, Franck, Pauline, Julie, Goge, Lolo et Mat pour leur soutien.

Enfin, j'en arrive aux remerciements qui me tiennent particulièrement à cœur, et pour cause, ceux que j'adresse à ma famille. Une famille atypique mais tellement extraordinaire. Je remercie donc mes parents ainsi que Annick et Alphonse pour l'amour et la confiance qu'ils m'ont accordé, pour m'avoir permis de réaliser mes études dans les meilleures conditions et pour leur soutien durant cette période qui a été bien plus longue que prévue! Ce travail vous ai donc dédié car votre contribution dans sa réussite est sans aucun doute, la plus importante. Je tiens également à souligner l'importance du Dr. Alphonse Iyamuremye dans ma décision de me lancer dans cette aventure et du Professeur des écoles, M. Gilbert Laxalde, pour les nombreuses re-lectures de ce manuscrit à la recherche de la moindre faute d'orthographe, bien que moindre ne soit pas approprié! Je tiens également à saluer mes (nombreux!) frères et sœurs Floriane, Jennifer, Tifenn, Gwénolé et Ewen pour leurs encouragements et leur affection. Je remercie également mon Grand-Père et ma Mamie et je veux honorer la mémoire de ma grand-mère, qui nous a quitté durant cette période de thèse.

Pour finir, je remercie bien entendu celle qui partage ma vie, Laëtitia, pour le réconfort et la stabilité qu'elle m'a apporté dans les moments difficiles et pour le soutien et l'amour qu'elle m'apporte depuis de nombreuses années.

Table des matières

Remerciements . . . 3
Liste des abréviations . . . 11

Introduction **17**

1 Produits lourds du pétrole **21**
1.1 Composition chimique des produits pétroliers . . . 22
1.1.1 Les hydrocarbures . . . 23
1.1.2 Les composés soufrés . . . 24
1.1.3 Les composés azotés . . . 25
1.1.4 Les composés oxygénés . . . 25
1.1.5 Les composés métalliques . . . 26
1.1.6 Le fractionnement SARA . . . 26
1.2 Raffinage du pétrole . . . 29
1.2.1 La distillation . . . 29
1.2.2 Le raffinage "conventionnel" . . . 30
1.2.3 Le raffinage des coupes lourdes . . . 31
1.3 Méthodes analytiques pour la caractérisation des produits lourds . . . 35
1.3.1 Les propriétés physico-chimiques globales . . . 36
1.3.2 Les répartitions massiques par familles chimiques . . . 37
1.3.3 Les analyses élémentaires . . . 38
1.3.4 Caractéristiques des méthodes analytiques . . . 39
1.4 Bilan . . . 43

2 Analyse rapide des produits lourds du pétrole : un bilan **45**
2.1 Techniques spectroscopiques . . . 47
2.1.1 La spectroscopie moyen infrarouge . . . 47
2.1.2 La spectroscopie proche infrarouge . . . 49
2.1.3 Comparaison des spectroscopies PIR et MIR . . . 51
2.2 Outils chimiométriques . . . 52
2.2.1 Prétraitements et domaines spectraux . . . 52

2.2.2 Méthodes d'étalonnage multivarié . . . 54
2.3 Détermination des propriétés des produits lourds . . . 55
2.3.1 Propriétés physico-chimiques globales . . . 55
2.3.2 Répartitions massiques par familles chimiques . . . 57
2.3.3 Analyses élémentaires . . . 57
2.4 Bilan . . . 58

3 Démarche analytique : matériels et méthodes associés 63
3.1 La démarche analytique . . . 63
3.2 La base de données . . . 67
3.2.1 Base d'échantillons . . . 67
3.2.2 Acquisition des spectres PIR . . . 74
3.2.3 Acquisition des spectres MIR . . . 75
3.3 Les algorithmes génétiques . . . 76
3.3.1 Principe général des algorithmes génétiques . . . 78
3.3.2 Avantages et limitations des AG . . . 82
3.3.3 Application des AG à la co-optimisation . . . 82
3.4 La fusion de données spectrales . . . 86
3.4.1 *Multiblock* PLS . . . 87
3.4.2 *Serial* PLS . . . 90
3.5 La comparaison de modèles . . . 92
3.5.1 Le Randomisation t-test . . . 93
3.5.2 Les méthodes de *bootstrap* . . . 96
3.6 Bilan . . . 99

4 Développement de modèles d'analyse multivariée 101
4.1 Exploration des données . . . 102
4.1.1 Présentation des caractéristiques de la base d'échantillons SARA . . . 102
4.1.2 Interprétation des spectres . . . 104
4.1.3 Présentation des bases spectrales . . . 108
4.1.4 Visualisation par ACP . . . 112
4.1.5 Bilan . . . 117
4.2 Optimisation simultanée du prétraitement et de la sélection de variables des modèles PIR par algorithmes génétiques . . . 118
4.2.1 Démarche de l'optimisation par AG . . . 119
4.2.2 Résultats et discussion . . . 125
4.2.3 Bilan . . . 135
4.3 Comparaison et fusion des spectroscopies MIR et PIR . . . 137
4.3.1 Démarche . . . 138
4.3.2 Performances des modèles . . . 139
4.3.3 Interprétation des résultats . . . 143
4.3.4 Bilan . . . 151
4.4 Conclusions . . . 153

5 Caractérisation globale des produits lourds par spectroscopie PIR **155**
5.1 Description général des modèles développés . . . 156
5.2 La détermination des propriétés physico-chimiques globales . . . 159
5.2.1 La densité . . . 159
5.3 La détermination des propriétés de répartition massique par familles chimiques . . . 162
5.3.1 Les performances des modèles SAR et asphaltènes C7 . . . 162
5.3.2 La teneur en carbone Conradson . . . 166
5.4 La prédiction des teneurs en éléments . . . 168
5.4.1 La teneur en carbones insaturés . . . 168
5.4.2 La teneur en hydrogène . . . 170
5.4.3 La teneur en azote . . . 172
5.4.4 La teneur en soufre . . . 174
5.5 Conclusions . . . 176

Conclusions et Perspectives **177**

Bibliographie **183**

A Techniques chimiométriques complémentaires **195**
A.1 Prétraitements mathématiques . . . 195
A.1.1 Dérivation . . . 195
A.1.2 Méthodes de correction de ligne de base . . . 197
A.1.3 Normalisation . . . 197
A.2 Méthodes de sélection des échantillons . . . 199
A.2.1 Sélection aléatoire . . . 199
A.2.2 Méthode de "Kennard et Stone" . . . 199
A.2.3 Méthode SPXY . . . 200
A.3 Régression PLS . . . 200
A.4 Validation croisée . . . 202
A.5 Critères statistiques . . . 204

B Résultats complémentaires **207**
B.1 Comparaison des spectroscopies MIR et PIR . . . 207
B.2 Caractérisation globale des produits lourds par spectroscopie PIR . . . 210
B.2.1 La viscosité . . . 210
B.2.2 La teneur en asphaltènes . . . 212
B.2.3 La teneur en carbone . . . 214

C Communications Scientifiques sur la période de la thèse **217**

Liste des abréviations

ACP :	Analyse en Composantes Principales
AG :	Algorithme génétique
ANN :	Réseaux de Neurones Artificiels
ATR :	*Attenuated Total Reflectance*
BiPLS :	*Backward Interval PLS*
DSV :	Distillat Sous-Vide
DTGS :	*Deuterated Triglycine Sulfate*
FCC :	*Fluid Catalytic Cracking*
IC :	Intervalle de confiance
MB-PLS :	*Multiblock-PLS*
MIR :	Moyen Infrarouge
MLR :	*Multi-Linear Regression*
MSC :	*Multiplicative Scatter Regression*
PCR :	*Principal Component Regression*
PIR :	Proche Infrarouge
PLS :	*Partial Least Squares*
RA :	Résidu Atmosphérique
RMN	Résonance magnétique nucléaire
RMSEC :	*Root Mean Square Error of Calibration*
RMSECV :	*Root Mean Square Error of Cross Validation*
RMSEP :	*Root Mean Square Error of Prediction*
RSV :	Résidu Sous-Vide
SARA :	Saturés, Aromatiques, résines et Asphaltènes
SNV :	*Standard Normal Variate*
S-PLS :	*Serial PLS*
WLSB :	*Weighted Least Square Baseline*

Table des figures

1.1 Composition d'un pétrole brut et relation point d'ébullition/masse molaire/structure . . 22
1.2 Illustrations des différents modèles de structure des asphaltènes : types continental et archipel 28
1.3 Distillation atmosphérique et sous vide d'un pétrole brut et exemples de coupes associées 30

2.1 Schéma de la propagation du faisceau infrarouge en ATR-IR 48

3.1 Démarche de développement d'une analyse multivariée 64
3.2 Méthodologie pour la constitution de la base d'échantillons 68
3.3 Gamme analytique de propriétés en fonction de la densité 71
3.4 Origine géographique des échantillons 73
3.5 Nombre d'échantillons provenant des différents procédés 73
3.6 Le spectromètre PIR 74
3.7 Étapes de la procédure d'optimisation par algorithmes génétiques 79
3.8 Opération génétiques pour la création de nouveaux individus 81
3.9 Représentation du codage des prétraitements et de la sélection de variables pour la co-optimisation par AG 83
3.10 Principe de la MB-PLS 88
3.11 Principe de la S-PLS 91
3.12 Procédure du randomisation t-test 95
3.13 Procédure *bootstrap* pour la génération d'un échantillon *bootstrap* 98

4.1 Gammes analytiques des teneurs en SARA couverte par les échantillons de la base . . . 103
4.2 Spectres MIR d'un DSV et d'un RA 105
4.3 Spectres PIR d'un DSV et d'un RA 106
4.4 Spectres MIR bruts des échantillons de la base 109
4.5 Spectres PIR bruts des échantillons de la base 110
4.6 Spectres MIR en dérivée 1$^{\text{ère}}$ sur le domaine 4000-3200 et 2700-1700 cm^{-1} 112
4.7 Spectres PIR en dérivée 1$^{\text{ère}}$ sur le domaine 9000-4000 cm^{-1} 113
4.8 Synthèse des résultats de l'ACP sur les spectres MIR 114

4.9 Synthèse des résultats de l'ACP sur les spectres PIR 116
4.10 Démarche de l'optimisation par AG pour chaque propriété 123
4.11 Le prétraitement *Weighted Least Square Baseline* (WLSB) 128
4.12 Interprétation des variables sélectionnées dans le cadre de la co-optimisation par AG . . 130
4.13 Coefficients de régression PLS des modèles PIR 144
4.14 Coefficients de régression PLS des modèles MIR 146
4.15 Interprétation du modèle MB-PLS de prédictions des teneurs en résines 149
4.16 Modèle MB-PLS et S-PLS de prédiction des teneurs en asphaltènes C7 151

5.1 Résultats pour le modèle de détermination des valeurs de la densité 161
5.2 Performances du modèle de détermination de la teneur en saturés 163
5.3 Résultats pour le modèle de détermination de la teneur en aromatiques 163
5.4 Résultats pour le modèle de détermination de la teneur en résines 164
5.5 Performances du modèle de détermination de la teneur en asphaltènes C7 165
5.6 Résultats pour le modèle de détermination de la teneur en carbone Conradson 167
5.7 Résultats pour le modèle de détermination de la teneur en carbones insaturés 169
5.8 Interprétation du modèle de prédiction de la teneur en hydrogène 171
5.9 Interprétation du modèle de prédiction de la teneur en azote 173
5.10 Interprétation du modèle de prédiction de la teneur en soufre 175

A.1 Effet de la dérivée sur les effets additifs et additifs plus multiplicatifs 196
A.2 Choix du nombre de composantes par validation croisée 204

B.1 Interprétation du modèle de prédiction de la teneur en viscosité 211
B.2 Résultats pour le modèle de détermination de la teneur en asphaltènes 213
B.3 Résultats pour le modèle de détermination de la teneur en carbone 215

Liste des tableaux

1.1 Structure de composés hydrocarbonés présents dans les produits pétroliers 24
1.2 Structure de composés soufrés présents dans les produits pétroliers 24
1.3 Structure de composés azotés présents dans les produits pétroliers 25
1.4 Structure de composés oxygénés présents dans les produits pétroliers 26
1.5 Structure de composés métalliques présents dans les produits pétroliers 26
1.6 Exemples de propriétés globales des coupes d'un pétrole brut (Moyen-orient) 31
1.7 Caractéristiques des différents procédés 35
1.8 Propriétés de caractérisation des produits lourds 36
1.9 Caractéristiques des méthodes analytiques des produits lourds 42

2.1 Résultats bibliographiques pour la détermination de la SARA 60
2.2 Résultats bibliographiques pour la détermination de la viscosité 61
2.3 Résultats bibliographiques pour la détermination des autres propriétés 62

3.1 Caractéristiques de la base d'échantillons 70
3.2 Les 32 prétraitements disponibles lors de la co-optimisation par algorithmes génétiques . 85

4.1 Gamme analytique en fonction de la coupe pétrolière 111
4.2 Paramètres des AG pour la co-optimisation 121
4.3 Présentation et comparaison des modèles PLS, P-GAPLS et Co-GAPLS pour la prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7 127
4.4 Localisation des blocs de variables sélectionnés par rapport aux bandes d'absorption connues 132
4.5 Présentation et comparaison des modèles PLS, MB-PLS et S-PLS pour la prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7 140

5.1 Résultats du développement de l'analyse multivariée des produits lourds par spectroscopie PIR 158

B.1 Présentation et comparaison des modèles PLS, MB-PLS et S-PLS pour la prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7 209

Introduction

Depuis le début de l'ère industrielle, le pétrole a été massivement utilisé comme combustible pour la production d'énergie. Les nombreuses crises environnementales et l'impact des rejets liés à son utilisation ont contribué à une prise de conscience sur la nécessité de diversifier les sources de production d'énergie. Néanmoins, les besoins énergétiques mondiaux ne cessent d'augmenter en raison de l'accroissement de la population et de l'émergence de pays en voie de développement tels que l'Inde et la Chine. Dans ce contexte, le pétrole demeure la ressource la plus consommée au niveau mondiale (34%) devant le charbon (30%), le gaz (24%) et l'électricité (12%) [1]. Faute d'être renouvelable à l'échelle des temps humains, une pénurie de pétrole est à craindre.

La modernisation des techniques exploratoires et d'extraction permet désormais d'exploiter des champs de pétrole jusqu'alors difficiles d'accès et d'accroître les ressources disponibles. Cependant, l'exploitation de champs de plus en plus profonds et de bruts non-conventionnels aboutit à la production de pétroles dont la fraction lourde est plus abondante. La valorisation de ces produits lourds n'est possible que si elle est accompagnée du développement de procédés de raffinage performants pour les convertir en produits légers (carburants) et réduire leurs teneurs en hétéroéléments et en métaux. Les recherches menées sur ces procédés ont pour but d'améliorer l'activité des catalyseurs et d'affiner les conditions opératoires afin d'optimiser les performances de ces procédés.

L'évaluation de l'impact de ses recherches sur les performances des procédés de valorisation passe principalement par la caractérisation des effluents et, notamment des produits

lourds. Cependant, du fait de leur nature et de leur complexité, l'analyse des fractions lourdes par les méthodes de référence est chronophage. Le nombre d'analyses disponibles pour le suivi des procédés est par conséquent restreint pour des raisons de coûts et de délais. Ceci entraîne un pilotage non-optimal des unités de raffinage. De plus, ces études sont réalisées sur des unités pilotes dont la capacité de production est faible par rapport aux unités en raffineries industrielles. La récupération du volume de produit nécessaire à la carcatérisation est donc également un point critique.

Pour répondre aux besoins analytiques du suivi des procédés de valorisation des produits lourds, nous proposons de développer une analyse spectroscopique, par définition rapide et nécessitant peu de volume d'échantillon. La spectroscopie infrarouge est une technique de choix pour la détermination indirecte des propriétés des produits lourds du pétrole. En effet, les bandes de vibration observées dans ce domaine sont très liées à la composition chimique des produits et des hydrocarbures en l'occurence. De plus, l'acquisition des spectres nécessite généralement peu de préparation de l'échantillon. Enfin, par le biais d'une analyse multivariée, il sera possible de prédire simultanément les propriétés d'intérêt d'un nouvel échantillon à partir de l'acquisition de son spectre.

Afin de garantir l'efficacité d'un modèle prédictif multivarié, plusieurs facteurs doivent être considérés. Tout d'abord, la fidélité des mesures de référence et de l'acquisition des spectres joue un rôle prépondérant. De plus, la base d'échantillons utilisée doit être représentative des produits qui seront analysés dans le futurs. Enfin, il est nécessaire de traiter les données spectrales afin de corriger ou d'éliminer les variations non-désirées, et ainsi de limiter l'introduction d'erreurs lors du calcul de l'équation d'étalonnage multivarié. Cette étape est sujet à une optimisation pour identifier les domaines spectraux d'intérêt et pour déterminer les méthodes de prétraitements mathématiques des spectres les plus adaptées.

L'objectif de cette thèse est de développer une analyse des produits lourds du pétrole par spectroscopie infrarouge capable de caractériser la plus grande partie des effluents des procédés de valorisation. Il en résulte que les échantillons analysés dans le futur seront très disparates en termes d'origines géographiques et de coupes (gamme de points d'ébullition). Ils seront également issus de procédés différents. De plus, comme nous l'avons mentionné, ces recherches menées pour l'amélioration des performances des procédés des produits

lourds consistent à tester de nouvelles conditions opératoires et de nouveaux catalyseurs. Contrairement aux produits finis, les échantillons analysés ne répondent pas, dans ce cas, aux spécifications et les gammes analytiques considérées sont alors très étendues.

Les produits lourds sont composés des molécules de plus hauts points d'ébullition et de plus grandes tailles. Leur composition est très complexe et polydispersée. L'obtention d'une base représentative de la diversité des produits lourds sera donc un des points importants de ce travail. De plus, ces échantillons sont opaques et très absorbants dans le domaine de l'infrarouge. De très faibles trajets optiques doivent alors être utilisés. Ces produits étant également très visqueux, leur échantillonnage est problématique et nécessite de les maintenir en température. Des protocoles expérimentaux spécifiques seront développés afin de limiter les variations de température pendant l'acquisition du spectre et, ainsi, réduire au maximum les erreurs de mesures spectrales. Enfin, la considération de produits de compositions chimiques très différentes et de gammes analytiques étendues peut soulever des problèmes lors du développement des modèles prédictifs multivariés tels que la présence de non-linéarités ou de groupes d'échantillons correspondant à des signatures, chimiques ou spectrales, spécifiques. Le cœur de ce travail consistera alors à optimiser le calcul des étalonnages multivariés afin de corriger ces problèmes. Il est envisagé de faire appel à des techniques chimiométriques peu communes telles que l'application d'un algorithme d'optimisation pour le choix des paramètres des modèles ou l'exploitation simultanée de différentes techniques spectroscopiques.

Ce manuscrit sera articulé autour de cinq chapitres. Le chapitre 1 permettra au lecteur de se familiariser au domaine pétrolier. Pour ce faire, nous présenterons la composition chimique des produits pétroliers, les différentes opérations de raffinage qu'ils peuvent subir et, enfin, les méthodes d'analyse pour leur caractérisation. Le chapitre 2 est un bilan des recherches menées pour l'analyse des produits lourds du pétrole. Les différentes techniques spectroscopiques et les méthodes d'analyse multivariée qui ont été employées dans la littérature seront exposées. Dans le but d'évaluer la faisabilité de cette approche, nous examinerons par la suite les résultats obtenus dans la littérature pour la détermination par spectroscopie infrarouge des différentes propriétés d'intérêts des produits lourds. Le chapitre 3 présentera la démarche mise en œuvre pour le développement de l'analyse mul-

tivariée des produits lourds du pétrole. Nous décrirons également le matériel employé et les méthodes appliquées. Le chapitre 4 concernera les travaux effectués pour le développement de l'analyse multivariée et, notamment pour l'optimisation des modèles dans le but d'améliorer leur pouvoir prédictif. Ces travaux ont été réalisés sur un nombre de propriétés restreints mais qui sont représentatives de l'ensemble des analyses d'intérêts. Pour finir, le chapitre 5 listera l'ensemble des propriétés qui ont fait l'objet du développement d'un étalonnage multivarié. Ces modèles ont néanmoins été développés à partir de méthodes chimiométriques classiques et n'ont pas fait l'objet d'une optimisation spécifique approfondie. L'objectif de ce chapitre est seulement de donner une vision d'ensemble de l'analyse rapide des produits lourds et du travail effectué au cours de cette thèse.

1 Produits lourds du pétrole

L'objectif de ce chapitre sera d'initier le lecteur au domaine pétrolier et d'introduire la terminologie propre à ce domaine. Le pétrole est un continuum de molécules dont la majorité sont des hydrocarbures. Ces hydrocarbures sont très hétérogènes en termes de structures. Des impuretés, telles que les hétéroéléments et les métaux, sont également présentes. Nous présenterons dans le détail la composition chimique des produits pétroliers.

La distillation permet de fractionner le pétrole brut en coupes pétrolières (essence, kérosène, gazole...) en fonction de la température d'ébullition des molécules qui le composent. Les produits lourds (ou coupes lourdes) correspondent aux composés dont la température d'ébullition est supérieure à 350°C. Ils correspondent à la partie la plus complexe et polydisperse du pétrole. Les produits lourds sont également les moins valorisés économiquement. Nous exposerons les différentes opérations de raffinage développées actuellement pour leur valorisation.

Afin d'effectuer le suivi des procédés de valorisation et d'évaluer leurs performances, il est nécessaire de caractériser les produits lourds. Cependant, leur analyse est difficile du fait de leur nature (opaque et visqueux) et de leur complexité. Les méthodes analytiques utilisées pour la caractérisation des produits lourds seront détaillées dans ce chapitre.

1.1 Composition chimique des produits pétroliers

Le pétrole est un mélange complexe majoritairement constitué d'hydrocarbures (93 à 99 %(m/m) [1]) 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. Il est composé d'un continuum de molécules hydrocarbonées pouvant comporter de quelques unités à plus d'une centaines d'atomes de carbone.

La Figure 1.1, représente la composition des pétroles bruts en fonction de la température d'ébullition et de la masse molaire. La masse molaire est estimée à partir du nombre d'atomes de carbone sur la base de la formule brute des paraffines C_nH_{2n+2} (alcanes).

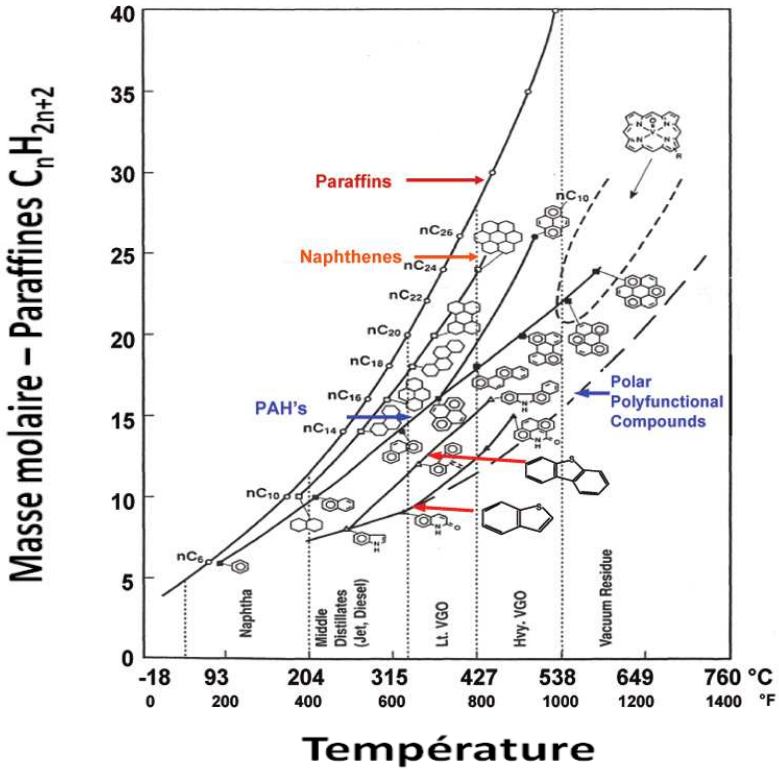


FIGURE 1.1 – Composition d'un pétrole brut et relation point d'ébullition/masse molaire/structure [78].

Cette figure illustre le fait que les matrices pétrolières sont très complexes et polydispersées. En effet, le nombre d'isomères augmente de manière exponentielle avec le nombre

1. %(m/m) : pourcentage massique

d'atomes de carbone. Par conséquent, plus les valeurs de températures d'ébullition et de masses molaires augmentent, plus les molécules présentes sont hétérogènes en termes de structures. Si on se place, par exemple, à la température d'ébullition de 427°C, le nombre d'atomes de carbone peut varier de 10 à plus de 25, et les structures des molécules correspondantes couvrent aussi bien les paraffines (alcanes) que les composés polycycliques.

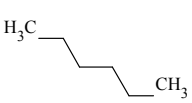
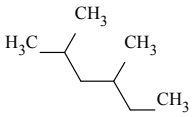
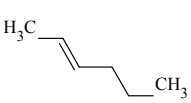
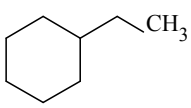
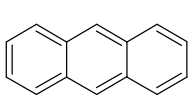
Les différents constituants du pétrole sont classés par familles chimiques : hydrocarbures, composés hétéroatomiques, composés métalliques... Cette partie a pour but de présenter ces différentes familles chimiques. Nous nous référons au livre *Pétrole brut, produits pétroliers, schémas de raffinage* de Wauquier *et al.* [106] qui est un ouvrage de référence dans le domaine pour la présentation des produits pétroliers.

1.1.1 Les hydrocarbures

Une distinction des hydrocarbures par familles chimiques peut-être effectuée en fonction du degré d'insaturations de la structure des molécules (Tableau 1.1) :

- Les **hydrocarbures aliphatiques saturés** : paraffines normales dites N-paraffines (alcanes linéaires) ou iso-paraffines (alcanes ramifiés).
- Les **hydrocarbures aliphatiques insaturés** : oléfines (alcènes) qui ne se rencontrent pas ou très peu dans le pétrole brut du fait de leur réactivité. Cependant, les oléfines peuvent être produites lors des procédés de raffinage et notamment lors de procédés de conversion des coupes lourdes.
- Les **hydrocarbures aliphatiques cycliques saturés** : naphtènes qui sont des cycles carbonés de 5 ou 6 atomes pouvant comporter un ou plusieurs cycles et des chaînes ramifiées.
- Les **hydrocarbures aromatiques** : 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 1.1 – Structure de composés hydrocarbonés présents dans les produits pétroliers

Familles	N-paraffines	Iso-paraffines	Oléfines	Naphtènes	Aromatiques
Formules	C_nH_{2n+2}	C_nH_{2n+2}	C_nH_{2n}	C_nH_{2n}	C_nH_{2n-8k}
Exemples					

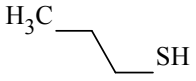
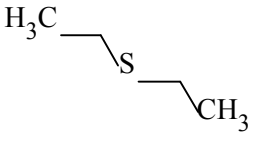
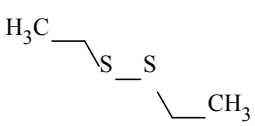
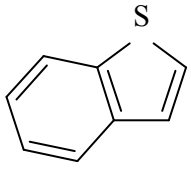
n : nombre d'atomes de carbones
k : nombre d'insaturations

1.1.2 Les composés soufrés

Le soufre est l'hétéroélément le plus répandu dans le pétrole. Sa teneur est fortement corrélée avec la densité de la matrice. Ainsi, les coupes lourdes en contiennent la majeure partie. Les composés soufrés présents dans les produits pétroliers appartiennent à différentes familles chimiques (Tableau 1.2) :

- Les **thiols ou mercaptans** : composés acides et corrosifs surtout présents dans les coupes légères, de formule brute $R - S - H$.
- Les **sulfures** : peu corrosifs et inodores du fait de leur faible volatilité, de formule brute $R - S - R'$ (ou polysulfures de formule brute $R - S - \cdots - S - R'$).
- Les **composés thiophèniques** qui présentent un caractère aromatique.

Tableau 1.2 – Structure de composés soufrés présents dans les produits pétroliers

Familles	Mercaptans	Sulfures	Disulfures	Benzothiophènes
Exemples				

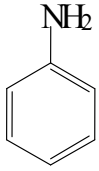
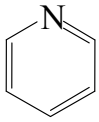
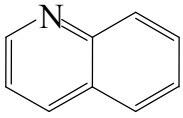
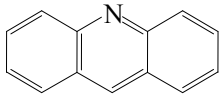
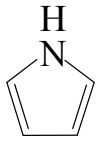
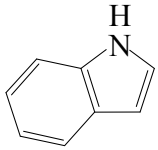
Les composés soufrés sont à l'origine de pollution atmosphérique (SO_2 et SO_3) et de la désactivation de certains catalyseurs utilisés notamment en procédés de raffinage ou dans les pots catalytiques. Les spécifications européennes régissant la teneur maximale

en soufre dans les carburants sont d'ailleurs régulièrement abaissées. En 2010, la teneur maximale en soufre dans les essences et les gazoles a été fixée à 10 ppm (EN228 et EN590).

1.1.3 Les composés azotés

Les composés azotés sont essentiellement présents dans les fractions lourdes, en plus faible quantité que les composés soufrés. Ils se distinguent essentiellement suivant leur caractère neutre ou basique (Tableau 1.3). Les composés azotés basiques et, dans une moindre mesure les composés azotés neutres, sont connus pour empoisonner les catalyseurs acides. Ils constituent alors un obstacle au raffinage des coupes lourdes.

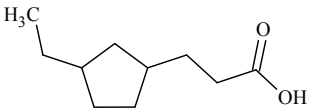
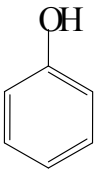
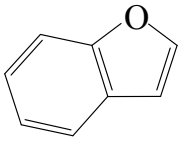
Tableau 1.3 – Structure de composés azotés présents dans les produits pétroliers

Familles	**Dérivés basiques**				**Dérivés neutres**	
	Aniline	Pyridine	Quinoléine	Acridine	Pyrrole	Indole
Exemples						

1.1.4 Les composés oxygénés

Parmi les composés oxygénés présents dans les produits pétroliers, on peut distinguer les acides carboxyliques naphténiques, les esters, les phénols, les furanes et les benzofuranes (Tableau 1.4). Bien qu'ils soient présents en faibles teneurs, principalement dans les coupes lourdes, les composés oxygénés possèdent un caractère acide qui est responsable de l'acidité globale des pétroles bruts et qui engendre des problèmes de corrosion.

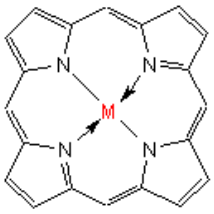
Tableau 1.4 – Structure de composés oxygénés présents dans les produits pétroliers

Familles	Acide naphténique	Phénol	Furane	Benzofurane
Exemples				

1.1.5 Les composés métalliques

Le nickel et le vanadium sont les métaux les plus répandus dans les produits pétroliers et sont principalement présents dans les produits lourds. Ils se trouvent dans des macros molécules dont les structures sont mal connues. Ils sont également présents dans certains composés plus petits de la famille des porphyrines. Dans cette structure, le motif est constitué par un ensemble de quatre cycles pyrroliques, le métal étant au centre de cet ensemble sous la forme Ni^{++} ou VO^{+} (Tableau 1.5). Bien qu'ils soient présents à très faibles teneurs, ils sont également des poisons pour les catalyseurs utilisés pour la conversion des coupes lourdes.

Tableau 1.5 – Structure de composés métalliques présents dans les produits pétroliers

Familles	Porphyrines
Exemple	

M=Ni^{2+} ou VO^{+}

1.1.6 Le fractionnement SARA

Compte tenu de la complexité des produits pétroliers lourds, il est d'usage de les séparer en fonction de leur polarité pour obtenir quatre fractions que sont les saturés, les aromatiques, les résines et, pour les produits les plus lourds, les asphaltènes (fractionne-

ment SARA).

La définition générale présente les asphaltènes comme la fraction insoluble d'une matrice pétrolière dans un solvant paraffinique (n-heptane ou n-pentane). A partir de la fraction soluble dans ce même solvant, appelée maltènes, les fractions saturées, aromatiques et résines sont séparées par chromatographie en phase liquide en utilisant un gradient de polarité obtenu par différents jeux de solvants. Cependant, les caractéristiques structurales ou chimiques des molécules qui composent les résines et les asphaltènes sont mal connues. En effet, ces fractions contiennent des molécules de grandes tailles (nombre d'atomes de carbone élevé). Le nombre d'isomères potentiellement présents est alors gigantesque (par exemple, 10^5 pour des molécules comprenant 20 atomes de carbones [10]). Par conséquent, leur analyse complète détaillée est très difficile voire impossible.

Les asphaltènes représentent la fraction contenant les composés les plus polaires et de plus hautes masses moléculaires. Ils contiennent une grande partie des hétéroéléments et des métaux présents dans les produits pétroliers. Compte tenu de cette complexité et de leur très grande hétérogénéité, il n'existe pas de motif structural unique et plusieurs modèles ont été avancés pour décrire leur structure [79].

Le modèle de Yen *et al.* [114] décrit le squelette carboné des asphaltènes caractérisé à l'état solide à l'aide des notions de molécules, de particules et d'agrégats. Il montre que l'unité de base est un feuillet polyaromatique péricondensé, comportant des hétérocycles, supportant des groupements fonctionnels et des substituants alkyls. On parle alors d'asphaltènes de type continental (Figure 1.2). De récents travaux décrivent les nano-agrégats d'asphaltènes comme une structure au cœur de forme cylindrique, contenant des noyaux polycycliques empilés. La taille de ce cylindre serait limitée par répulsion stérique due à des chaînes périphériques [39]. Un autre modèle [81] décrit les molécules d'asphaltènes comme étant des noyaux aromatiques reliés par des chaînes alkyles hétéroatomiques ou non. On parle alors d'asphaltènes de type archipel (Figure 1.2).

Un repliement de la construction place les feuillets ou les cycles, selon le modèle, en empilement dont la cohésion est assurée par les électrons π des doubles liaisons des cycles benzéniques. Ces caractéristiques confèrent aux asphaltènes des propriétés telles que l'auto-association, la floculation, la sédimentation et la précipitation dans des envi-

ronnements organiques ou lors des différents traitements que le pétrole peut subir. Ces propriétés sont à l'origine de la formation de coke [2] et du bouchage des unités de raffinage.

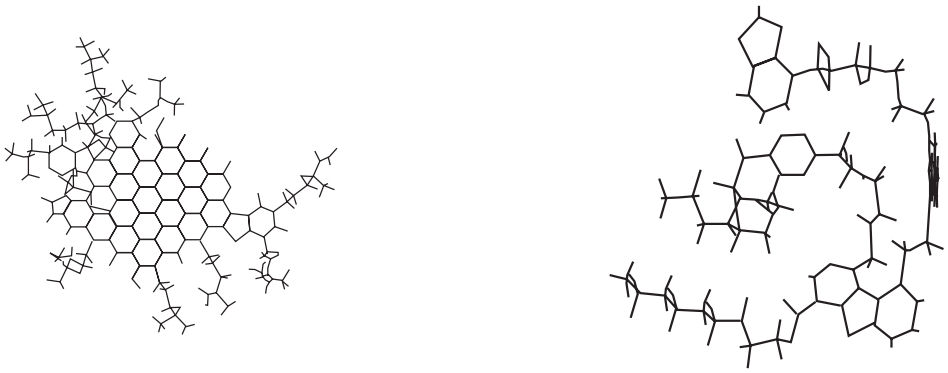


FIGURE 1.2 – Illustrations des différents modèles de structure des asphaltènes : types continental (à gauche) et archipel (à droite) [79]

Les résines contiennent des structures aromatiques (polycondensées ou non) dont le nombre de cycles est supérieur à 6, ainsi qu'une partie des hétéroéléments et des métaux. Elles jouent un rôle essentiel dans la stabilité du pétrole en prévenant la démixtion des asphaltènes. La fraction résine est considérée comme moins aromatique que la fraction asphaltène. Cependant, la polarité globale des résines n'est qu'un peu plus faible que celle des asphaltènes [96]. De ce fait, la frontière entre les résines et les asphaltènes est très difficile à déterminer. En effet, cette frontière est fonction de la matrice pétrolière considérée et, il est envisageable qu'une même molécule puisse se retrouver, suivant l'échantillon, dans la fraction résine ou asphaltène.

La fraction aromatique est formée de molécules contenant jusqu'à 5 ou 6 cycles aromatiques (condensés ou non) et pouvant être alkylés. Une partie des aromatiques soufrés est également contenue dans cette famille. Des hétéroéléments comme l'oxygène et l'azote peuvent aussi être présents en quantités limitées. En revanche, on ne trouve pas de métaux dans cette famille.

Les saturés sont composés d'hydrocarbures saturés de type normal et iso paraffines ainsi que de cycles naphténiques. Cette famille ne contient quasiment pas d'hétéroéléments.

2. Le coke de pétrole se présente sous forme de solide noir, se compose essentiellement de carbone, de très peu d'hydrogène et d'importantes quantités de polluants (Partie 1.2.3.1)

1.2 Raffinage du pétrole

Le raffinage du pétrole désigne l'ensemble des traitements et transformations visant à produire le maximum de produits à hautes valeurs commerciales à partir du pétrole brut, telles que les bases carburants (essences, kérosènes, gazoles et fuels), les intermédiaires pour la pétrochimie, les plastiques...

1.2.1 La distillation

La distillation est une étape préliminaire au raffinage des produits pétroliers. Elle permet de fractionner le pétrole brut afin d'obtenir différentes coupes pétrolières en fonction de la température d'ébullition (Figure 1.3). Tout d'abord, la distillation atmosphérique permet de séparer les coupes gaz (<35°C, noté 35°C$^-$), essence (35-175°C), kérosène (175-235°C) et gazole (235-350°C). La partie du produit pétrolier qui n'a pas été distillée lors de cette opération est appelé le résidu atmosphérique (RA) et est composée des molécules dont le point d'ébullition est supérieur à 350°C, noté 350°C$^+$. Le résidu atmosphérique peut ensuite être séparé, par distillation sous vide, en deux autres coupes pétrolières : le distillat sous vide dit DSV (350-550°C) et le résidu sous vide dit RSV (>550°C, noté 550°C$^+$). Cette opération est effectuée sous vide afin d'éviter le craquage des molécules qui se produit au delà de 400°C. Le craquage est le phénomène qui correspond à la *cassure* d'une molécule complexe (de grande taille) en éléments plus petits. Nous pouvons également noter qu'une coupe pétrolière provenant directement du pétrole brut, c'est à dire qui n'a subit aucune opération de raffinage, sera dite de "distillation directe".

Ce que l'on appelle communément les produits lourds du pétrole sont définis comme les composés dont le point d'ébullition est supérieur à 350°C, noté 350°C$^+$. Ils comprennent donc les coupes RA, DSV et RSV. Les produits lourds sont constitués des composés de plus hautes températures d'ébullition et de plus hautes masses moléculaires. Ils sont donc les produits les plus complexes du pétrole et sont très hétérogènes en termes de structure. De plus, ils sont très opaques (noirs) et très visqueux (solide à température ambiante).

Les coupes pétrolières sont donc caractérisées par une gamme de température d'ébullition et par un nombre d'atomes de carbone. Leur proportion dans le pétrole brut et

leur composition dépendent de l'origine géographique du pétrole et de la coupe considérée (Tableau 1.6). En effet, on remarque que l'évolution de la densité et de la proportion des impuretés (hétéroéléments et métaux) est fonction de la coupe pétrolière. Un pétrole brut sera ainsi d'autant plus valorisable économiquement qu'il sera riche en coupes légères et pauvre en impuretés, car il nécessitera moins d'étapes de raffinage.

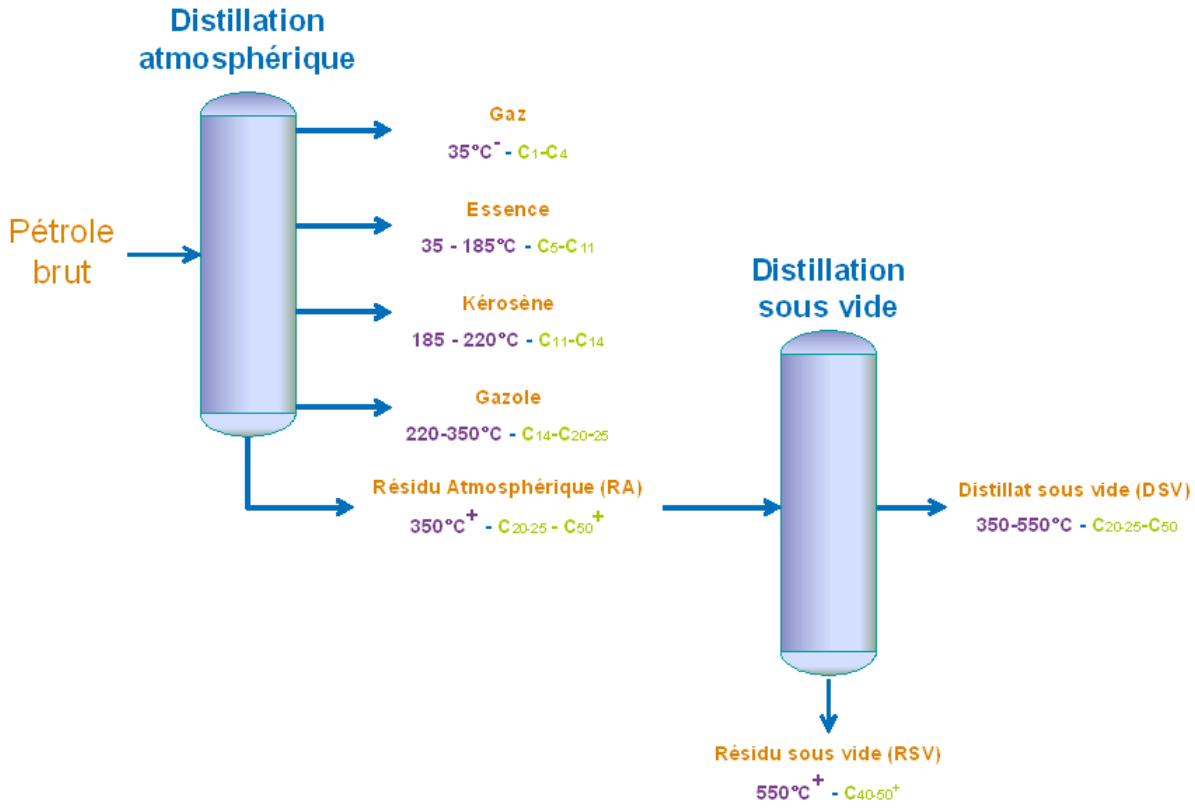


FIGURE 1.3 – Distillation atmosphérique et sous vide d'un pétrole brut et exemples de coupes associées

1.2.2 Le raffinage "conventionnel"

Les produits commerciaux sont soumis à des spécifications qui garantissent leurs performances et qui limitent leur impact sur l'environnement (teneur en soufre, en aromatiques polycycliques...). Les différentes opérations de raffinage ont pour but de transformer ces produits afin d'obtenir la plus grande quantité possible de bases carburants valorisables répondant aux spécifications. Ainsi, les essences seront transformées pour

améliorer leur indice d'octane [3]. Les gazoles subiront des opérations visant à améliorer leur indice de cétane [4] et à réduire leur teneur en soufre. Les kérosènes seront utilisés pour les transports aériens après amélioration, notamment, de leur tenue à froid. Enfin, les produits lourds produisent également des dérivés commerciaux tels que les fiouls lourds utilisés comme carburant pour les centrales thermiques ou pour les navires, les huiles lubrifiantes qui sont des DSV déparaffinés et le bitume produit à partir du RSV.

Tableau 1.6 – Exemples de propriétés globales des coupes d'un pétrole brut (Moyen-orient) [106]

Coupes	**Gaz**	**Essence**	**Kérosène**	**Gazole**	**DSV**	**RSV**
Nombre d'atomes de carbone	$C_1 - C_4$	$C_4 - C_{10}$	$C_{10} - C_{14}$	$C_{14} - C_{25}$	$C_{25} - C_{55}$	C_{35}^{+}
Intervalle d'ébullition (°C)	<0	0-180	180-230	230-375	375-600	600^{+}
Rendement sur brut (%(m/m))	1,37	17,72	6,74	24,37	23,50	26,30
Densité (d_4^{15})	0,654	0,742	0,793	0,851	0,935	1,037
Soufre (%(m/m))	0,003	0,035	0,150	1,400	3,800	5,000
Azote (ppm)	-	-	-	-	≈ 1000	≈ 2000

1.2.3 Le raffinage des coupes lourdes

La demande pétrolière mondiale s'oriente de plus en plus vers les carburants et, il est acquis que la production de pétroles bruts conventionnels ne pourra pas répondre à cette demande croissante dans un futur proche. Ainsi, l'industrie pétrolière doit faire face à la production de pétrole brut dont la fraction lourde est de plus en plus importante.

Dans ce contexte, de nouveaux développements sont attendus pour la conversion des coupes lourdes en coupes valorisables. De plus, ces nouvelles applications permettraient d'étendre le raffinage aux pétroles bruts non-conventionnels tels que les sables bitumeux (Canada) ou des bruts dont la fraction lourde est très importante (Vénézuela). En effet,

3. l'indice d'octane mesure la résistance d'un carburant à l'auto-allumage
4. L'indice de cétane sert à apprécier l'aptitude à l'auto-inflammation d'un gazole

cette alternative est très attractive car les ressources de pétroles bruts non-conventionnels sont deux fois supérieures aux ressources de bruts conventionnels.

Dans cette partie, les différents procédés pour la valorisation des produits lourds du pétrole sont brièvement présentés. Ces procédés ont deux principaux objectifs :

- le traitement pour réduire les teneurs en polluants (hétéroéléments et métaux).
- le craquage (ou conversion) qui visent à craquer les liaisons afin de réduire le nombre d'atomes par molécules (convertir les molécules de grandes tailles en molécules plus petites)

Ces procédés peuvent se distinguer en trois classes : les procédés thermiques (viscoréduction ou cokéfaction), les procédés catalytiques sans apport d'hydrogène et les procédés catalytiques avec apport d'hydrogène.

1.2.3.1 Les procédés thermiques

La viscoréduction est un procédé de craquage thermique modéré (450-500°C) sans catalyseur qui permet la conversion des RA ou des RSV en produits exploitables (carburants). Les effluents de viscoréduction sont souvent instables, riches en hétéroéléments et en oléfines. Une étape d'hydrotraitement (Partie 1.2.3.3) est souvent nécessaire pour les incorporer aux produits finis. A noter que le distillat de viscoréduction est parfois employé comme charge pour le craquage catalytique fluide (Partie 1.2.3.2).

La cokéfaction est un procédé thermique à haute température (environ 500°C pendant plusieurs heures) qui permet la valorisation des résidus sous vide. La cokéfaction produit des effluents liquides et solides. Les effluents liquides doivent subir un hydrotraitement avant d'être réinsérés dans les opérations de raffinage "conventionnel", car ils sont très riches en aromatiques, oléfines, soufre et azote. Les effluents solides se présentent sous forme de coke de pétrole. Le coke de pétrole peut être utilisé comme combustible ou calciné pour former un composé à plus de 98 %(m/m)de carbone qui sert à la fabrication d'électrodes. Ces électrodes sont ensuite utilisées dans l'industrie de la métallurgie notamment pour la production d'aluminium ou d'acier.

1.2.3.2 Les procédés catalytiques sans apport d'hydrogène

Le craquage catalytique fluide ou *Fluid Catalytic Cracking* (FCC) est le principal procédé catalytique sans apport d'hydrogène. Il est très important pour la production d'essences et dans une moindre mesure de gazoles. Procédé très flexible, il traite des produits telles que les DSV, qui peuvent être issus de distillation directe ou des procédés thermiques, et des RSV ayant une faible teneur en métaux. L'opération se déroule en phase gazeuse, à basse pression et à des températures comprises entre 500 et 540°C. Les produits de conversion du craquage catalytique sont largement oléfiniques pour les fractions légères et fortement aromatiques pour les fractions lourdes, avec un rendement en gaz important.

1.2.3.3 Les procédés catalytiques avec apport d'hydrogène

Les procédés catalytiques avec apport d'hydrogène ont pour but de convertir les fractions les plus lourdes du pétrole en produits valorisables. Ces procédés font l'objet de nombreux développements. Dans ce manuscrit, seuls les trois principaux procédés catalytiques avec apport d'hydrogène développés à IFP Energies Nouvelles sont abordés. Le Tableau 1.7, résume les caractéristiques de ces trois procédés.

- Le **procédé d'hydrocraquage** (HCK) des DSV en lit fixe
- Le **procédé Hyvahl**^{TM} : hydrotraitement des RA et RSV en lit fixe
- Le **procédé H-Oil®** : hydroconversion des RA et RSV en lit bouillonnant

Le procédé d'hydrocraquage traite les coupes lourdes exemptes d'asphaltènes telles que les DSV qui peuvent provenir de distillation directe ou être issus de procédés de conversion de produits plus lourds types RA ou RSV (viscoréduction, cokéfaction, hydrotraitement ou hydroconversion). Ce procédé se déroule en deux étapes. Une première étape est consacrée à l'hydrotraitement afin d'éliminer les impuretés (hétéroéléments, métaux...) et notamment l'azote. L'hydrodésazotation est une opération indispensable en amont de l'étape d'hydrocraquage car elle permet de préserver les catalyseurs acides des effets d'empoisonnement. En effet, ces catalyseurs sont très sensibles à la présence des composés azotés basiques qui s'adsorbent et abaissent très fortement leur activité. La conversion d'hydrocraquage proprement dite est effectuée dans un deuxième temps. C'est un procédé qui permet une conversion élevée des DSV (70-100 %(m/m)). De plus,

la qualité des effluents d'hydrocraquage est très bonne (très stables et faibles teneurs en impuretés).

Le procédé Hyvahl^{TM} est un procédé très efficace pour l'amélioration de la qualité des RA ou RSV. Il est bien adapté au traitement des charges contenant moins de 100-120 ppm de métaux (nickel et vanadium). Ce procédé comporte plusieurs réacteurs en série. Les premiers réacteurs sont dédiés à l'hydrodémétallisation (HDM) et à la conversion, tandis que les suivants sont consacrés à l'hydrodésulfuration (HDS) et au raffinage. Le premier réacteur, qui est principalement consacré à l'HDM, peut être décuplé afin de permettre le traitement de charges à teneurs plus élevées en métaux (jusqu'à 400 ppm). Ainsi, lorsque le catalyseur d'un de ces réacteurs est désactivé, il est isolé de l'ensemble des autres réacteurs pendant la maintenance. Suivant les charges et les objectifs, deux options sont envisagées. Dans l'option conversion maximum, le niveau de conversion du RA peut atteindre 60-70 %(m/m). Dans l'option HDS, la conversion est sensiblement réduite. Dans les deux cas, les niveaux d'HDS et d'HDM sont supérieurs à 90 %(m/m). Les résidus hydrotraités sont stables et de bonne qualité (peu d'hétéroéléments et de métaux).

Le procédé H-Oil® était initialement conçu pour la conversion profonde des résidus en produits légers, puis a été appliqué au traitement des résidus à teneurs particulièrement élevées en impuretés. La section réactionnelle comprend un ou plusieurs réacteurs à lit bouillonnant disposés en série. La mise en œuvre du procédé H-Oil® est plus difficile que celle du procédé Hyvahl^{TM} en raison des opérations de renouvellement de catalyseur qui s'effectuent en ligne, à température élevée et à haute pression. La consommation de catalyseurs est élevée et la qualité des produits est inférieure à celle provenant des autres procédés catalytiques avec apport d'hydrogène. En particulier, à forte conversion, le résidu non converti est en limite de stabilité. Cependant, les niveaux de conversion du RSV peuvent être très élevés et atteindre 80 %(m/m).

Tableau 1.7 – Caractéristiques des différents procédés

Type de procédés	**Hyvahl ®**	**H-Oil ®**	**HCK**
État du catalyseur	lit fixe	lit bouillonnant	lit fixe
Tolérance aux impuretés	faible	Moyenne	faible
Teneur max en Ni+V (ppm)	120-400	>700	0
Conversion max du résidu (%w/w)	60-70	80	70-100
Qualité de l'effluent	Bonne	Moyenne	Très bonne
Stabilité de l'effluent	oui	limite	oui
Mise en œuvre	Bonne	Difficile	Bonne

1.3 Méthodes analytiques pour la caractérisation des produits lourds

La caractérisation des produits pétroliers est primordiale pour, d'une part, déterminer s'ils correspondent aux critères de qualité des produits finis et, d'autre part, pour évaluer l'impact des procédés sur les caractéristiques physico-chimiques des produits en vue de leur optimisation. Nous pouvons noter que, dans l'industrie pétrolière, les méthodes d'analyses sont normalisées. Ces normes peuvent être internationales telles que les normes ASTM ("American Society for Testing and Materials") ou ISO (International Organization for Standardization), française (NF) ou interne (IFP Energies Nouvelles).

Les analyses réalisées pour la caractérisation des produits lourds du pétrole peuvent être regroupées en trois grands types : les propriétés physico-chimiques globales, les répartitions massiques par familles chimiques et les analyses élémentaires (Tableau 1.8).

Cette partie a pour but d'introduire ces différentes analyses et de présenter leur intérêt pour la caractérisation des produits lourds. Le Tableau 1.9, situé à la fin de cette partie, résume pour chaque propriété la durée d'analyse, le volume consommé, le domaine d'application et l'incertitude de mesure associée.

Tableau 1.8 – Propriétés de caractérisation des produits lourds

Propriétés physico-chimiques globales	Densité Indice de réfraction Viscosité (Pa.s ou cSt) Point d'écoulement (°C)
Répartitions massiques par familles chimiques	Fractionnement SAR(A) (%(m/m)) Carbone Conradson (CCR) (%(m/m)) Teneur en asphaltènes (%(m/m))
Analyses élémentaires	Teneur en carbone (%(m/m)) Teneur en carbones insaturés (%(m/m)) Teneur en hydrogène (%(m/m)) Teneur en soufre (%(m/m)) Teneur en azote (%(m/m)) Teneur en nickel (%(m/m)) Teneur en vanadium (%(m/m))

1.3.1 Les propriétés physico-chimiques globales

Les propriétés macroscopiques les plus utilisées sont la masse volumique ou densité, l'indice de réfraction et la viscosité.

La mesure de la densité est très utilisée dans l'industrie pétrolière car elle donne une indication rapide, fiable et reproductible de la qualité d'une coupe pétrolière. En effet, plus la densité sera faible, plus la coupe aura un caractère paraffinique. Inversement, plus elle sera élevée, plus le caractère aromatique sera prédominant. Ainsi, la densité est fortement corrélée à un grand nombre de propriétés telles que les teneurs en carbones insaturés, en hydrogène, en SARA ... La masse volumique est déterminée par deux méthodes en fonction de l'état du produit à 70°C (liquide ou solide). Dans les deux cas, le résultat est exprimé en densité d_4^{15}, qui est le rapport de la masse volumique du produit à 15°C par rapport à celle de l'eau mesurée à 4°C.

L'indice de réfraction est l'une des déterminations les plus précises qui puisse être conduite sur un produit pétrolier. En effet, la reproductibilité sur les produits les plus visqueux est de 6.10^{-4}. Il peut donc servir à la différenciation de deux produits très proches. En revanche, l'indice de réfraction des produits les plus lourds ne peut-être déterminé du fait de leur opacité.

La viscosité désigne la capacité d'un fluide à s'écouler à une température donnée.

Beaucoup de produits pétroliers sont utilisés comme lubrifiants et le bon fonctionnement des appareils dépend de l'utilisation d'une huile d'une viscosité appropriée. C'est une des propriétés les plus difficiles à déterminer et aucune méthode n'est vraiment satisfaisante pour tous les intervalles de température et de viscosité. En pratique, la viscosité des produits pétroliers se détermine par deux méthodes : la viscosité dynamique (ou absolue) et la viscosité cinématique. L'analyse de viscosité dynamique est utilisée pour les produits les plus lourds et se mesure en Pascal-seconde (Pa.s). La viscosité cinématique est définie comme le rapport entre la viscosité dynamique et la masse volumique [106]. Elle s'exprime en $mm^2.s^{-1}$ ou en centistockes (cSt) [5].

Le point d'écoulement est la température la plus basse à laquelle un produit pétrolier peut encore s'écouler dans des conditions normalisées. Sa connaissance peut être très importante pour caractériser les propriétés à froid des produits pétroliers. Cette propriété est essentielle dans l'industrie automobile ou aéronautique (carburants, lubrifiants) mais également dans les problématiques de transport des produits pétroliers (pipelines).

1.3.2 Les répartitions massiques par familles chimiques

Du fait de la complexité de la composition chimiques des produits pétroliers, l'analyse détaillée est difficile voire impossible. Afin d'obtenir des informations plus spécifiques sur ces échantillons, ils sont fractionnés en fonction de leur aromaticité (SARA et Asphaltènes C7) ou après combustion (Carbone Conradson).

Le principe de la méthode de la SARA (Partie 1.1.6) est de déterminer le pourcentage massique de chaque fraction dans l'échantillon. Pour cela, l'échantillon est injecté en chromatographie à phase liquide qui va séparer les fractions maltènes (saturés, aromatiques et résines). Leurs teneurs respectives, exprimées en pourcentage massique, sont ensuite déterminées par pesée. Dans le cas des produits pétroliers qui contiennent plus 1 %(m/m)d'asphaltènes, une séparation préalable des asphaltènes par un solvant paraffinique est réalisée. Nous pouvons souligner que cette méthode est très longue et fastidieuse. En effet, le fractionnement SARA pour un échantillon nécessite plus d'une journée d'analyse et 8 heures de temps technicien.

5. 1 cSt= 1 $mm^2.s^{-1}$

La détermination du résidu de carbone ou carbone Conradson est très utilisée industriellement pour estimer la capacité d'un produit à former du coke. En effet, il est un promoteur de la désactivation des catalyseurs des procédés de conversion. Le résidu de carbone est déterminé en chauffant l'échantillon d'une température inférieure à 100°C jusqu'à 500°C avec un gradient de 10 à 15 °C.min^{-1} puis en le maintenant à cette température pendant 15 min. Les fractions volatiles formées pendant la combustion sont entraînées par un courant d'azote. Le résidu restant, qui correspond au carbone Conradson, est pesé.

La teneur en asphaltènes est une propriété importante. En effet, les asphaltènes ont tendance à s'agréger et à floculer ce qui entraîne le bouchage des unités de raffinage ou des problèmes de stockage. La teneur en asphaltènes C7 est le pourcentage en masse des constituants insolubles dans l'heptane mais solubles dans le toluène chaud. Cette analyse s'applique aux teneurs en asphaltènes comprises entre 0,5 et 30 %(m/m). Elle diffère de l'analyse de la SARA car le protocole de récupération des asphaltènes y est différent. En particulier, les composés insolubles dans le toluène à chaud sont compris dans les asphaltènes de la SARA.

1.3.3 Les analyses élémentaires

Afin d'obtenir des informations complémentaires sur la composition des produits pétroliers, l'analyse des éléments présents (C, H, O, N, S, Ni, V ...) est effectuée. La présence d'hétéroéléments ou de composés organométalliques affecte la qualité du produit pétrolier et augmente les contraintes lors des procédés de valorisation des produits lourds. En effet, ils sont, avec la formation de coke, les principaux promoteurs de la désactivation des catalyseurs. Connaître leurs teneurs dans les produits pétroliers est donc indispensable afin d'adapter le procédé et de suivre leur élimination.

La détermination du nickel et du vanadium est réalisée par fluorescence X. Le principe de la méthode consiste à irradier l'échantillon par un rayonnement X primaire, de mesurer les intensités des rayons X secondaires caractéristiques de l'atome concerné et de les comparer à celles obtenues sur des références. Cette méthode s'applique à des teneurs en nickel de 2 à 600 ppm et en vanadium de 2 à 1300 ppm.

1.3.4 Caractéristiques des méthodes analytiques

La notion d'incertitude de mesure est rappelée dans cette partie en s'appuyant sur les normes ISO 3534-1 [6] et sur les publications de l'IUPAC [7]. Les caractéristiques des méthodes utilisées pour l'analyse des produits lourds sont ensuite indiquées dans le Tableau 1.9.

Les mesures réalisées pour estimer une grandeur physique ou chimique ne sont pas exactes. En effet, elles sont entachées d'erreurs de mesure. Le calcul d'incertitude permet d'évaluer les erreurs qui se produisent lors de ces mesures. Afin d'estimer cette incertitude de mesure, deux grandeurs sont à prendre en compte : la justesse et la fidélité.

La justesse est définie comme l'écart entre la valeur mesurée expérimentalement et la valeur vraie. Elle dépend uniquement des erreurs systématiques telles que les défauts d'étalonnage. Le biais peut-être utilisé comme estimateur de la justesse d'une méthode analytique. En effet, plus le biais est grand, plus la justesse est faible. Si x désigne le résultat analytique et que l'on peut disposer de la valeur x_0, la valeur vraie ou la valeur certifiée de l'échantillon de référence, le biais Δ est donné par $\Delta = x - x_0$. Cependant, le biais est difficile à estimer car il est souvent impossible de disposer d'étalons dans la même matrice que celle analysée. Par conséquent, la justesse d'une méthode est estimée par comparaisons inter-laboratoires ou en réalisant l'analyse à partir de plusieurs techniques.

La fidélité est définie comme étant l'accord entre des résultats indépendants obtenus sous des conditions stipulées. En d'autres termes, la fidélité est l'aptitude de la méthode à donner des résultats les plus proches possibles lors d'analyses répétées d'un même échantillon. Le terme "résultats d'essais indépendants" signifie des résultats obtenus d'une façon non influencée par un résultat précédent sur le même matériau d'essai ou similaire. La fidélité dépend uniquement de la distribution des erreurs aléatoires. Par conséquent, lorsque la fidélité est mesurée pour une valeur donnée, elle n'a théoriquement aucune relation avec la valeur vraie. La mesure de fidélité est exprimée en terme d'infidélité et est calculée à partir de l'écart-type des résultats d'essais. Les mesures quantitatives de la fidélité dépendent de façon critique des conditions stipulées. Les conditions de répétabilité et de

6. (*Statistique - Vocabulaire et symboles - Partie 1 : Termes statistiques généraux et termes utilisés en calcul des probabilités*)

7. (*International Union of Pure and Applied Chemistry*)

reproductibilité sont des ensembles particuliers de conditions extrêmes.

La répétabilité est mesurée sous des conditions où les résultats d'essais indépendants sont obtenus par la même méthode sur des individus d'essais identiques dans le même laboratoire, par le même opérateur, utilisant le même équipement et pendant un court intervalle de temps.

La reproductibilité est mesurée sous des conditions où les résultats d'essais sont obtenus par la même méthode sur des individus d'essais identiques dans différents laboratoires, avec différents opérateurs et utilisant des équipements différents.

La répétabilité ou la reproductibilité sont estimées à partir de l'écart-type (Equation 1.1).

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1.1}$$

avec x_i : i^{ème} valeur obtenue sur une série de n mesures d'un échantillon

$\bar{x}$: valeur moyenne, sur la série de n mesures

n : nombre de mesures

Si l'on suppose que la distribution des résultats est normale, une mesure x_i est ensuite encadrée par l'intervalle de confiance (Equation 1.2).

$$x_i \pm t^{\nu}_{1-\alpha/2} \times s \tag{1.2}$$

avec $t^{\nu}_{1-\alpha}$: le fractile de la loi de Student

$\nu = n$ - 1 : le degré de liberté où n représente le nombre de mesures qui a servi au calcul de s

α : le niveau de risque associé (le plus souvent, $\alpha = 0{,}05$)

s : écart type de répétabilité ou de reproductibilité

L'incertitude globale d'une mesure cumule donc le biais et la fidélité. Les méthodes analytiques utilisées dans le domaine pétrolier sont normalisées. Ces normes définissent l'appareillage à utiliser ainsi que le protocole expérimental à suivre. Elles fixent également

les domaines d'applications des méthodes analytiques et fournissent une estimation de la fidélité (répétabilité, reproductibilité ou intervalle de confiance de la méthode). Du fait de la complexité des produits pétroliers, il est impossible de disposer d'étalons identiques à leur matrice. Par conséquent, la mesure du biais n'est pas réalisable. Le problème de la mesure de justesse est détourné par le fait que les laboratoires sont soumis à des essais circulaires.

Nous pouvons noter que la caractérisation d'un produit lourd par l'ensemble des analyses présentées précédemment nécessite un temps d'analyse d'environ 16 heures, dont 8 heures pour la SARA et un volume d'échantillon de 150 cm^3 (Tableau 1.9). A ce volume de produit, il faut ajouter le volume de liquide total nécessaire à la distillation pour obtenir la ou les coupes à analyser. Le liquide total est soit le pétrole brut, soit l'effluent que l'on récupère en fin de procédé et qui contient les différentes coupes pétrolières. De manière générale, pour obtenir un volume analytique suffisant pour chaque coupe, 5 litres de liquide total sont nécessaires.

Tableau 1.9 – Caractéristiques des méthodes analytiques des produits lourds

Propriété	**Norme**	**Domaine d'application[a]**	**Temps (h)**	**Volume (cm^3)**	**Fidélité**
Densité	ASTM D4052	0,68< y < 0,97	0,25	30	$r = 5.10^{-4}$
	NF T66-007	y> 0,97	0,5	20	$r = 2{,}6628.10^{-3}$
Indice de réfraction	ASTM D1218	y<1,60	0,4	10	$R = 6.10^{-5}$
Viscosité dynamique (Pa.s)	ASTM D3236-88	y<200	1,6	20	R = 0.254× y
Viscosité cinématique (cSt)	ISO 3104	produits liquides	1	60	R = 0,04×(y+8)
Le point d'écoulement (°C)	NF T60-105	∞	0,8	9	R = 0,04×(y+8)
Saturés %(m/m)		5-60			R = 3
Aromatiques %(m/m)	ASTM D2007	30-65	8	10	R = 3
Résines %(m/m)	ou ASTM D2549	3-40			R = 3
Asphaltènes %(m/m)		0-30			r = 0.14 × y
Carbone Conradson %(m/m)	ASTM D189	0,1 à 30,0	0,3	10	$R = 0{,}2451 \times y^{\frac{2}{3}}$
Asphaltènes C7 %(m/m)	NF T60-115	0,5 à 30, 0	1,4	30	r = 0.2 + 0.2 × y
Carbone %(m/m)		y>30			$IC = \pm 0.40$
Hydrogène %(m/m)		>1			$IC = \pm 0.20$
Soufre %(m/m)	externe	y < 1			$IC = \pm 0.10$
		1 <y < 10			$IC = \pm 0.20$
Azote (ppm)		500-10000			$IC = 300$
Oxygène (ppm)		>200			$IC = \pm 100$
Carbones insaturés %(m/m)	ASTM D5292-99	y > 1	3	2	$R = 2$
Nickel (ppm)	IFP 9422	2-600	0.5	0.5	$r = 0,008 \times y + 0,6$
Vanadium (ppm)		2-1300			$r = 0,001 \times y + 0,7$

[a] : exprimé dans l'unité de la propriété
y : valeur de la propriété
r : répétabilité (r)
R : reproductibilité (R)
IC : intervalle de confiance

1.4 Bilan

Nous avons vu que le pétrole est un continuum de molécules et que les fractions lourdes représentent la partie la plus complexe des produits pétroliers. En effet, les coupes lourdes sont composées de molécules de grandes tailles, très polaires, riches en hétéroéléments et en métaux.

Le contexte énergétique actuel pousse néanmoins l'industrie pétrolière à orienter ses recherches vers l'exploitation des produits lourds du pétrole. Ces produits font donc l'objet de nombreux travaux visant à optimiser leur conversion en produits valorisables et à limiter leur impact sur l'environnement.

L'optimisation de ces procédés nécessite la caractérisation des fractions lourdes. Les analyses de référence actuellement utilisées pour leur caractérisation sont cependant chronophages, coûteuses et consommatrices de produits. Le nombre d'analyses disponibles pour le suivi des procédés est par conséquent restreint par les coûts et les délais. Une analyse rapide est donc nécessaire pour fournir une analyse de suivi efficace et pour faire face à la demande analytique croissante soutenue par les recherches menées sur les procédés de valorisation. Dans le chapitre suivant, un bilan sera réalisé sur les travaux réalisés dans la littérature pour la caractérisation des produits lourds par une méthode rapide.

2

Analyse rapide des produits lourds du pétrole : un bilan

Dans le chapitre précédent, nous avons mentionné qu'il est nécessaire de développer une analyse rapide des produits lourds du pétrole afin de réduire les coûts et les délais pour leur caractérisation et, ainsi, proposer un contrôle efficace de leurs procédés de valorisation.

Pour réduire le temps de caractérisation d'un échantillon, il existe principalement deux approches : l'automisation des méthodes analytiques ou l'exploitation des propriétés physiques des produits en tant que source d'information [12]. Cependant, le gain obtenu par le biais de l'automisation des méthodes est limité par le temps de d'analyse. En effet, en chromatographie en phase gazeuse par exemple, l'installation d'un passeur automatique permettra d'analyser un grand nombre d'échantillons sans intervention de l'expérimentateur. Cependant, le temps de rétention reste inchangé ce qui limite le gain de temps. Les méthodes d'analyses rapides reposant sur l'exploitation des propriétés physiques des produits sont essentiellement des méthodes spectroscopiques. L'avantage de cette approche, par rapport à l'automisation des méthodes analytiques, est que le temps d'analyse est court et qu'il est possible de déterminer plusieurs propriétés simultanément à partir de l'acquisition d'un spectre [19].

Dans la littérature, les spectroscopies moyen infrarouge (MIR) et proche infrarouge (PIR) sont les techniques les plus utilisées pour l'analyse rapide des produits lourds du pétrole. Ceci peut s'expliquer par le fait que l'acquisition des spectres MIR et PIR est très rapide, peu consommatrice de produit et nécessite généralement peu de préparation de

l'échantillon. De plus, les spectroscopies MIR et PIR sondent les vibrations des liaisons organiques. Ainsi, l'information contenue dans les spectres MIR et PIR est très riche sur la composition des produits pétroliers [13, 54].

Cependant, l'analyse rapide à partir de spectres MIR et PIR est une méthode indirecte. En effet, à part dans quelques cas simples, il n'est pas possible de relier directement l'intensité des bandes aux valeurs des propriétés. Il est donc nécessaire de développer un étalonnage multivarié afin d'établir un modèle prédictif pour chaque propriété [77]. L'analyse multivariée a pour but d'extraire l'information spectrale pertinente pour la description de la propriété considérée.

Les produits lourds du pétrole, opaques et visqueux, sont difficiles à manipuler. De plus, les produits lourds sont très absorbants ce qui peut poser des problèmes d'échantillonnage en spectroscopies MIR et PIR. Dans la première partie de ce chapitre, nous nous focaliserons donc sur les modes d'échantillonnage et les protocoles expérimentaux (température d'acquisition, préparation de l'échantillon...) qui ont été utilisés pour leur analyse.

Le développement d'un étalonnage multivarié nécessite de faire appel à des méthodes chimiométriques. La deuxième partie aura donc pour but de lister les différentes méthodes qui ont été utilisées pour le développement de modèles de prédiction des propriétés des produits lourds.

Enfin, la troisième partie aura pour but d'évaluer le potentiel de l'analyse rapide des produits lourds par spectroscopie MIR et PIR. Pour cela, nous examinerons les travaux réalisés dans la littérature pour la prédiction de chaque propriété d'intérêt des produits lourds, dans le but d'évaluer la faisabilité de leur détermination par une analyse multivariée. Ainsi, cette étude prend en compte les travaux portant sur tous types de produits lourds : les dérivés commerciaux (huiles lubrifiantes, bitumes...), les produits de distillation directe, les effluents des procédés de conversion ainsi que les pétroles bruts qui contiennent également des produits lourds.

Afin de faciliter la lecture, les différents travaux réalisés pour l'analyse rapide des produits lourds par spectroscopies MIR et PIR sont résumés dans trois tableaux à la fin de ce chapitre. Pour chaque modèle développé dans la littérature, ces tableaux indiquent le

nombre et le type de produits pétroliers analysés, l'étendue de la propriété considérée et la technique spectroscopique utilisée. Le prétraitement, le domaine spectral et la technique de régression appliqués pour le développement du modèle prédictif sont également renseignés. Enfin, les critères statistiques pour l'évaluation des erreurs de prédiction sont mentionnés. La définition et la manière dont sont obtenues ces valeurs de RMSEC, RMSECV et RMSEP sont rappelées dans les Parties A.4 et A.5. Le Tableau 2.1, résume les travaux pour la détermination des teneurs en SARA, le Tableau 2.2, les modèles pour la prédiction de la viscosité et le Tableau 2.3, les études pour les autres propriétés.

2.1 Techniques spectroscopiques

Bien que les spectroscopies MIR et PIR soient les techniques spectroscopiques les plus utilisées, l'analyse rapide des produits lourds du pétrole a également été réalisée à partir d'autres techniques spectroscopiques telles que la spectroscopie Raman [24], la RMN du proton [29, 45, 82], du carbone [29] et la spectroscopie de fluorescence [89]. Cependant, ces méthodes sont peu employées pour le développement de modèles prédictifs. En effet, le coût de l'instrumentation peut-être très élevé et ces méthodes sont pour la plupart difficiles à intégrer en industrie, notamment, dans le cadre d'une analyse en ligne. De plus, elles nécessitent souvent une préparation de l'échantillon en amont de l'analyse, une dilution notamment.

Les différents modes d'échantillonnage qui ont été utilisés dans la littérature en spectroscopie MIR et PIR pour l'analyse des produits lourds du pétrole sont présentés dans cette partie. En effet, la nature de ces produits, qui sont opaques et visqueux, nécessite l'utilisation d'appareillages spécifiques.

2.1.1 La spectroscopie moyen infrarouge

La spectroscopie MIR correspond à la région du spectre électromagnétique qui s'étend de 4000 à 400 cm^{-1} (2500 - 25000 nm). Les bandes d'absorption observées dans cette région sont principalement dues aux vibrations fondamentales des molécules. Généralement, les bandes observées sont bien résolues et relativement spécifiques [87]. Ainsi, il est possible

d'attribuer la plupart des bandes à un groupement chimique spécifique. Cette plage spectrale est donc particulièrement bien adaptée à l'identification des composés organiques et à l'étude de la conformation des molécules [12].

Cependant, les coefficients d'absorption dans le MIR sont très importants. Il est ainsi nécessaire de travailler avec des trajets optiques très faibles afin d'éviter la saturation du signal. Il peut donc être très difficile de procéder à l'acquisition des spectres des échantillons en mode transmission. C'est pourquoi, la réflexion total atténuée (ATR pour "Attenuated Total reflection") est très utilisée. Le principe de l'analyse par ATR est le suivant [87] : l'échantillon, d'indice de réfraction n_2, est déposé sur un cristal, d'indice de réfraction n_1, avec $n_1 > n_2$. Quand l'angle d'incidence est supérieur à l'angle critique θ [1], le faisceau subit alors une réflexion totale à l'intérieur du cristal (Figure 2.1). En réalité, le phénomène est perturbée par l'existence d'une onde progressive transversale appelée onde évanescente. Comme le montre la Figure 2.1, cette onde se propage dans l'échantillon et comme l'échantillon absorbe une partie de l'énergie lumineuse, la réflexion totale est en fait atténuée.

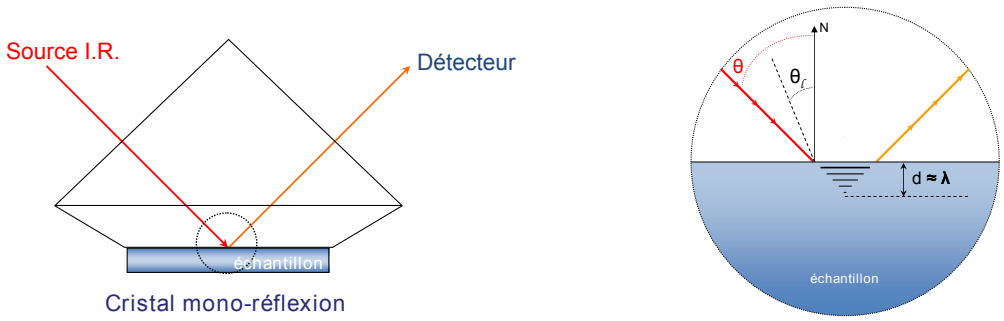


FIGURE 2.1 – Schéma de la propagation du faisceau infrarouge en mono-réflexion total atténuée (à gauche) et focus sur le phénomène d'onde évanescente (à droite)

Le mode d'échantillonnage ATR est très adapté à l'analyse des produits lourds du pétrole car le trajet optique est très faible. Ce mode permet alors l'acquisition de spectres MIR de produits lourds en évitant la saturation du signal, malgré la forte absorption de ces échantillons. Le protocole expérimental est de plus facilité car il suffit de déposer le produit sur le cristal. Il en résulte que l'ATR est le seul mode d'échantillonnage utilisé en

1. L'angle critique est donné par la loi de Snell-Descartes : $\sin(\theta) = \frac{n_2}{n_1}$

spectroscopie MIR pour l'analyse des produits lourds dans la littérature.

Du fait de la forte viscosité des produits lourds, l'acquisition des spectres en température semble nécessaire. En effet, la mise en température facilite l'échantillonnage de ces produits, le nettoyage du cristal et permet d'assurer un meilleur contact entre l'échantillon et le cristal. Ainsi, HONGFU *et al.* [54] ont analysé des RA et des RSV par ATR-IR. Les spectres ATR-IR ont été enregistrés à 60°C. Pour cela, ils ont préalablement chauffés les échantillons et la température est maintenu à 60°C ± 0.5° à l'aide d'un régulateur de température.

ASKE *et al.* [4] et HANNISDAL *et al.* [51] ont quant à eux déterminé les teneurs en fractions SARA dans les pétroles bruts par ATR-IR. Le problème d'échantillonnage est moins critique pour les pétroles bruts qui sont beaucoup moins visqueux. Les auteurs ont donc procédé à l'acquisition des spectres ATR-IR des pétroles bruts à température ambiante.

2.1.2 La spectroscopie proche infrarouge

Le domaine du PIR s'étend entre 12 500 et 4000 cm^{-1} (800 - 2500 nm). Dans cette partie du spectre électromagnétique, ce sont les vibrations harmoniques et les vibrations de combinaison des molécules qui sont observées. Dans la littérature, la spectroscopie PIR est la technique la plus utilisée pour la caractérisation des produits lourds. Ceci peut s'expliquer par le fait que les avantages de la spectroscopie PIR sont multiples.

Tout d'abord, l'intensité des bandes harmoniques et de combinaisons est de 10 à 10000 fois moins importante que celle des vibrations fondamentales [19]. Ainsi, il est possible de travailler avec des trajets optiques plus grands qu'en spectroscopie MIR et, par conséquent, d'analyser en mode transmission des produits très absorbants tels que les produits lourds. De plus, l'instrumentation de la spectroscopie PIR a de nombreux avantages tels que la grande émissivité des sources, la grande sensibilité des détecteurs et un rapport signal sur bruit élevé [84]. C'est pourquoi la spectroscopie PIR est très adaptée pour l'analyse en ligne en raffinerie [23]. Dans ce domaine, les trois modes d'échantillonnages sont employés dans la littérature : la transmission, la réflexion diffuse et la transflexion.

En mode transmission les trajets optiques utilisés dans la littérature sont de 200 μm

[13] ou 500 μm [24, 52, 86]. Afin d'échantillonner les produits dans ces cellules de très faibles épaisseurs, différents protocoles ont été utilisés.

Blanco *et al.* [13] ont analysé des bitumes par spectroscopie PIR. Les bitumes, obtenus à partir du RSV, sont les produits les plus lourds et, par conséquent les plus absorbants. Le trajet optique le plus faible (200 μm) a donc été choisi afin d'éviter la saturation du signal. L'échantillonnage de ces produits a été réalisé de la façon suivante : le bitume, contenu dans une boite de Pétri, est liquéfié à l'aide d'une lampe infrarouge puis aspiré dans la cellule de trajet optique de 200 μm. La cellule est ensuite maintenue à température ambiante pendant 20 minutes. Blanco *et al.* ont jugé que les spectres obtenus présentaient un signal répétable et un rapport signal/bruit satisfaisant.

Chung et Ku [24] ont procédé à l'acquisition des spectres PIR de RA en mode transmission avec un trajet optique de 500 μm. Pour cela, le RA est chauffé afin de le transférer dans la cellule. La température de l'échantillon est ensuite maintenue à 60°C pendant l'acquisition du spectre.

Enfin, Hidajat et Chong [52] et Pasquini et Bueno [86] ont analysé des pétroles bruts avec un trajet optique de 500 μm. Les pétroles bruts sont beaucoup moins visqueux que les RA ou les bitumes. Par conséquent, ils ont procédé à l'acquisition des spectres à température ambiante. On peut noter qu'ils mentionnent dans leur publication que la température de la pièce est régulée.

La transflexion a également été utilisée pour l'analyse de pétroles bruts [4, 51]. La transflexion met en jeu à la fois la transmission et la réflexion. La mesure de l'intensité de la lumière traversant l'échantillon est réalisée du même côté que la source lumineuse. Un miroir est placé derrière l'échantillon afin de réfléchir le signal. Le trajet optique est alors doubler dans l'échantillon. Ceci peut contribuer à accroître la précision pour des niveaux d'absorption chimique faibles. Aske *et al.* [4] ont utilisé un trajet optique total apparent de 200 μm et l'acquisition des spectres a été réalisée à 25°C (température ambiante). Hannisdal *et al.* [51] ont fait l'acquisition des spectres à température ambiante. Cependant, l'échantillon est préalablement chauffé pour l'homogénéiser. Ils ont utilisé un trajet optique total apparent de 320 μm.

La réflexion diffuse a été utilisée par Lima *et al.* [69, 70] pour l'analyse d'huiles lubri-

fiantes et de bitumes. Ils indiquent dans leur publication que la température de la pièce est maintenue aux alentours de 22°C pendant l'acquisition. Blanco *et al.* [13] ont comparé la réflexion diffuse à la transmission. Cependant, l'acquisition des spectres en mode réflexion n'est pas décrite car les spectres mesurés conduisaient à un signal faible et non répétable.

2.1.3 Comparaison des spectroscopies PIR et MIR

Les spectroscopies PIR et MIR sont les techniques les plus utilisées pour l'analyse rapide des produits lourds du pétrole. Ce paragraphe tente de déterminer, à partir des travaux réalisés dans la littérature, si l'une des deux techniques a une capacité analytique supérieure pour la prédiction des propriétés des produits lourds.

Tout d'abord, les travaux de différentes équipes sont difficilement comparables du fait de la disparité de l'instrumentation utilisée, des produits analysés et des gammes analytiques considérées. Toutefois, trois études de comparaison des spectroscopies MIR et PIR pour la caractérisation des produits lourds ont été menées dans la littérature.

Chung *et al.* [24] ont comparé ces deux techniques pour l'analyse de RA. Cependant, ils ont jugé que la qualité des spectres MIR obtenus en mode ATR était insuffisante. Par conséquent, aucun résultat n'est donné dans leur publication sur l'étalonnage multivarié à partir des spectres MIR.

Aske *et al.* [4] et Hannisdal *et al.* [51] ont développé une analyse multivariée pour la prédiction des teneurs en SARA dans les pétroles bruts à partir de spectres PIR et MIR. Les résultats obtenus par Aske *et al.* semblent indiquer que l'erreur de prédiction est plus faible à partir des spectres MIR pour les teneurs en saturés, aromatiques et résines. Pour les teneurs en asphaltènes, ils ont obtenus de meilleurs résultats à partir des spectres PIR. Hannisdal *et al.* aboutissent aux mêmes conclusions pour les teneurs en saturés, en résines et en asphaltènes. En revanche, pour la prédiction de la teneur en aromatiques, l'analyse développée à partir des spectres PIR semble plus performante.

Ces deux études montrent que les spectroscopies PIR et MIR peuvent être complémentaires pour la caractérisation des produits lourds. En effet, la capacité analytique de ces techniques dépend de la propriété considérée. Le nombre d'études étant très restreint,

aucune conclusion ne peut-être avancée à partir de la littérature sur la capacité analytique de ces deux techniques. Une étude comparative de ces deux spectroscopies pour la caractérisation des produits lourds du pétrole sera alors menée.

2.2 Outils chimiométriques

Dans cette partie, un bilan des techniques chimiométriques utilisées dans la littérature pour le développement de modèles d'analyse multivariée à partir des spectres MIR et PIR est effectué. Nous présenterons les prétraitements de spectres, le choix du domaine spectral et les algorithmes de régression appliqués. Nous ne détaillerons pas chacun des modèles développés mais essayerons de définir s'il existe des analogies entre les méthodes chimiométriques utilisées dans les différentes publications..

2.2.1 Prétraitements et domaines spectraux

L'objectif d'un prétraitement est d'éliminer les variations spectrales qui ne sont pas reliées à la propriété considérée. Ces variations spectrales peuvent être aléatoires (bruit) ou dues à l'instrumentation (non-linéarité du détecteur dans certaine gamme, variation du trajet optique) ou à des interférences physico-chimiques (diffusion). Les principaux prétraitements utilisés en chimiométrie sont décrits en Annexe A.1.

La plupart des modèles ont été développés en utilisant les spectres PIR et MIR dérivés. Les méthodes de dérivation ont la capacité de corriger à la fois les effets additifs (déplacements verticaux de la ligne de base ou "offset") et multiplicatifs (déplacements verticaux de la ligne de base en fonction de la longueur d'onde) qui peuvent apparaître dans les spectres [88]. Les spectres dérivés sont le plus souvent obtenus par la méthode de Savitzky-Golay [94]. Cette technique est décrite en Annexe A.1.1. Le calcul de la dérivée des spectres par la méthode de Savitzky-Golay nécessite de fixer trois paramètres : le degré de la dérivée, le degré du polynôme et la taille de la fenêtre Ces paramètres sont optimisés pour chaque propriété modélisée. Ils sont donc propres à l'étude menée et à la technique spectroscopique considérée.

Nous pouvons également noter que Long *et al.* [72] n'ont pas utilisé de prétraitement

pour développer leurs modèles de prédiction à partir des spectres PIR. Ils justifient leur choix par la stabilité de leur spectromètre. Ce choix est très discutable car, comme nous l'avons évoqué précedemment, les interférences présentes dans le spectre peuvent être provoquées par d'autres phénomènes que la mesure spectroscopique. Enfin, Hidajat et Chong [52] ont utilisé le prétraitement *Multiplicative Scatter Correction* (MSC). Ce prétraitement est généralement bien adapté pour la correction des problèmes de diffusion ou de variations du trajet optique [44, 76].

L'optimisation du domaine spectral ou l'utilisation d'une méthode de sélection de variables lors du développement d'une analyse multivariée peut également amener à une amélioration du pouvoir prédictif des modèles [97].

Dans les travaux réalisés pour la détermination des propriétés des produits lourds, beaucoup d'équipes ont procédé à une optimisation du domaine spectral en fonction de la propriété à prédire. Le choix *a priori* des domaines spectraux est justifié par la connaissance de la propriété et de l'information spectrale disponible. Les justifications de tous les auteurs, et pour toutes les propriétés, ne sont pas reprises ici dans le détail. Pour illustrer, nous avons pris l'exemple des travaux de Blanco *et al.* [15] pour la détermination de la SARA. Ils expliquent que l'information nécessaire pour déterminer les teneurs en saturés, aromatiques et résines se trouvent dans des régions spectrales spécifiques : les vibrations harmoniques et de combinaisons des liaisons C-H. Pour les saturés, Blanco *et al.* utilisent donc le domaine 6060 - 5560 et 4480 - 4020 cm^{-1}. Pour les aromatiques et les résines, ils étendent le domaine à 6060 - 5560 et 4760 - 4020 cm^{-1}, qui contient la combinaison des vibrations d'élongation des liaisons =C-H et C=C. Enfin, Blanco *et al.* ont développé le modèle de prédiction des teneurs en asphaltènes sur le spectre entier (9000 - 4020 cm^{-1}). Ils justifient ce choix par le fait que, même si les asphaltènes absorbent principalement dans la gamme des nombres d'onde élevés (9000-6500cm^{-1}), les chaînes hydrocarbonées qui relient les noyaux polyaromatiques absorbent également dans la région des faibles nombres d'onde (6500-4000cm^{-1}).

Cet exemple illustre la nécessité de sélectionner les variables pertinentes du spectres en fonction de la propriété à déterminer. Il a cependant été constaté qu'aucune technique de sélection de variables n'a été utilisée.

2.2.2 Méthodes d'étalonnage multivarié

Les méthodes linéaires les plus répandues pour l'étalonnage multivarié à partir de données spectrales sont la régression linéaire multiple (MLR pour *Multi-Linear Regression*) [17, 75], la régression en composantes principales (PCR pour Principal Component Regression) [9] et la régression par les moindres carrées partiels (PLS pour *Partial Least Square regression*) [112].

L'objectif de ces méthodes est d'établir une relation linéaire entre la matrice des spectres ($\mathbf{X}$) et le vecteur des valeurs de référence ($\mathbf{y}$). Parmi ces méthodes, la PLS fait office de référence. En effet, contrairement à la MLR, la PLS et la PCR comportent l'avantage de gérer les cas où les variables sont très corrélées entre elles et que le nombre de variables (longueurs d'onde) est beaucoup plus grand que le nombre d'observations (nombre d'échantillons) [100]. Pour cela, ces méthodes réduisent le nombre de variables des spectres en un nombre d'axes très inférieur à la taille de la matrice initiale. La PCR se déroule en deux étapes : tout d'abord, le nombre de variables est réduit par une analyse en composantes principales (ACP) puis une régression MLR est réalisée. La principale différence entre les régressions PLS et PCR réside dans le fait que lors de la régression PLS, la direction de ces nouveaux axes est directement définie dans le but de maximiser la covariance entre les variables du spectre ($\mathbf{X}$) et la propriété ($\mathbf{y}$) [43]. L'algorithme PLS est décrit dans en Annexe A.3.

La régression PLS est la technique la plus employée pour le développement de modèles de prédiction des propriétés des produits lourds. Cependant, les réseaux de neurones artificiels [21, 50] ont également été appliqués pour la détermination de la densité, du point d'écoulement et de la viscosité par Blanco *et al.* [14]. Pasquini et Bueno [86] ont également utilisé cette méthode pour la prédiction du degré API [2]. Les réseaux de neurones artificiels sont une méthode d'apprentissage qui peut s'avérer très utile pour pallier des problèmes de non-linéarités. Le principe est d'entraîner les réseaux de neurones afin qu'ils "apprennent" ce qui se passe dans le procédé que l'on traite sans connaître les lois physiques et chimiques qui gouvernent le système. Blanco *et al.* [14] ont obtenu des résultats similaires par PLS et par réseaux de neurones artificiels pour la prédiction de la densité.

2. API= $\frac{141.5}{densité\ (60°F/60°F)} - 131.5$

Pasquini et Bueno [86] ont obtenu de meilleurs résultats par PLS que par réseaux de neurones artificiels pour la détermination du degré API. Ces deux exemples semblent montrer que l'introduction de non-linéarité pour la détermination de la densité ne semble pas nécessaire. En revanche, Blanco *et al.* [15] ont mis en évidence une non-linéarité lors du développement du modèle de prédiction de la viscosité. Les auteurs ont attribué cette non-linéarité au fait que les produits lourds du pétrole sont des fluides "non Newtonien". La vitesse de déformation de ces produits n'est donc pas directement proportionnelle à la force qu'on leur applique, ce qui peut expliquer une relation non-linéaire avec certaines de leurs propriétés physico-chimiques et, notamment la viscosité. Afin de gérer cette non-linéarité, les auteurs ont comparé deux approches pour l'établissement de l'équation d'étalonnage : l'utilisation des réseaux de neurones et le développement d'un modèle prédictif à partir du logarithme des valeurs de viscosité. Ils ont obtenu un meilleur pouvoir prédictif en calculant un modèle PLS sur le logarithme des valeurs de viscosité qu'en développant un modèle sur les valeurs "brutes" par réseaux de neurones artificiels. Ce résultat peut s'expliquer par le fait que, malgré leur fort potentiel, les réseaux de neurones artificiels sont délicats à mettre en œuvre et qu'ils sont moins performants dans les cas où le nombre de variables est très élevé.

2.3 Détermination des propriétés des produits lourds

Dans cette partie, nous nous focaliserons sur les propriétés des produits lourds qui ont fait l'objet d'un développement de modèle prédictif. Le but est d'essayer de définir la faisabilité de la détermination de chaque propriété par une analyse par spectroscopie infrarouge. Ces travaux sont néanmoins difficilement comparable entre eux du fait de la disparité des produits, des gammes analytiques, des techniques spectroscopiques et des outils chimiométriques considérés.

2.3.1 Propriétés physico-chimiques globales

La détermination de la densité par spectroscopies PIR et MIR a fait l'objet de nombreuses publications [14, 15, 24, 45, 52, 54, 69, 93]. Les étendues considérées de densité

sont très différentes car cette propriété est très dépendante de la coupe analysée. Par exemple, Lima *et al.* [69] et Blanco et *et al.* [14, 15] ont respectivement analysé des huiles lubrifiantes et des bitumes dont les densités variées de 0,814 à 0,917 et de 1,021 à 1,091.

La prédiction de la viscosité a été abordée par spectroscopie PIR [14, 15, 70] et MIR [54]. Blanco *et al.* [14, 15] ont déterminé la viscosité des bitumes à 135 °C sur une gamme analytique très étendue (208-1056 cSt). Ils ont introduit des non-linéarités dans le modèle en utilisant différentes approches (Partie 2.2). Lima et Leite [70] et Hongfu *et al.* [54] ont respectivement analysé la viscosité de résidus hydrocraqués et de bitumes. Leurs modèles amènent à des erreurs de prédictions plus faibles mais sur des bases dont l'étendue en viscosité est beaucoup plus restreinte : 226 cSt d'étendue à 135°C pour une moyenne de 207 cSt et de 11,6 à 57 cSt à 100°C respectivement.

Un modèle de prédiction du point d'écoulement sur une étendue de -37 à -15°C a été développé par Chung et Ku [25]. Ils ont réalisé cette étude dans le cadre de la mise en ligne d'un analyseur PIR sur une unité de raffinage d'huiles de bases. Ils ont montré que, du fait de la faible reproductibilité de l'analyse de référence, l'analyse par spectroscopie PIR était beaucoup plus fidèle.

La détermination de l'indice de réfraction n'est pas mentionnée ici car aucune référence n'aborde sa prédiction par une analyse rapide. Ceci peut s'expliquer par la rapidité et la grande fidélité de la méthode de référence (Partie 1.3.4) et par le fait que cette méthode ne s'applique pas aux produits les plus lourds.

2.3.2 Répartitions massiques par familles chimiques

Les travaux menés dans la littérature montrent la faisabilité de la détermination de la SARA par spectroscopies PIR [4, 15, 51, 93] et MIR [4, 51, 54]. Cependant, il est difficile de tirer des conclusions sur les performances obtenues. En effet, les étendues de propriétés et les erreurs de prédiction sont très différentes selon les études.

Blanco *et al.* [15] et Long *et al.* [72] ont travaillé sur la modélisation de la teneur en asphaltènes C7 par spectroscopie PIR. Les prédictions obtenues par Blanco *et al.* [15] sont toutes dans l'intervalle de confiance de la méthode de référence sur une gamme de teneur en asphaltènes C7 allant de 9,9 à 26,8 %(m/m). Long *et al.* [72] ont travaillé sur la modélisation de la teneur en asphaltènes lors du procédé d'extraction du bitume contenu dans les sables bitumeux. Après l'extraction par un solvant aliphatique, ils obtiennent deux phases. La première contient le solvant, de l'eau et du bitume avec une teneur en asphaltènes comprise entre 0 et 20 %(m/m). La seconde phase est uniquement composée de bitumes avec une teneur en asphaltènes comprise entre 20 et 100 %(m/m). Long *et al.* [72] ont ensuite développé un modèle prédictif pour chacune des deux phase.

Hongfu *et al.* [54] ont déterminé la teneur en carbone Conradson par spectroscopie MIR sur une gamme analytique allant de 3 à 10 %(m/m). Ils ont obtenu une erreur de prédiction très faible. Satya *et al.* [93] et Gilbert *et al.* [45] ont également obtenu des résultats satisfaisants par spectroscopie PIR sur une gamme de teneurs en carbone Conradson allant respectivement de 0,1 à 15,7 %(m/m) et de 5 à 6 %(m/m).

2.3.3 Analyses élémentaires

Hongfu *et al.* [54] et Satya *et al.* [93] ont travaillé respectivement sur la teneur en hydrogène par spectroscopie MIR et sur le rapport H/C par spectroscopie PIR. Aucune référence ne fait état de la détermination de la teneur en carbone. Cependant, la prédiction des teneurs en hydrogène et en carbone est, *a priori*, réalisable. En effet, ces éléments sont les plus abondants dans les produits pétroliers et la plupart des absorptions dans les spectres de vibration des produits pétroliers sont dues aux liaisons comprenant des atomes de carbone et d'hydrogène.

Lima et Leite [70] ont montré que la différenciation du degré d'insaturation des carbones présents dans les bitumes par spectroscopie PIR est possible. En effet, les teneurs en carbones paraffiniques, naphténiques et aromatiques ont été déterminées par une analyse multivariée. L'erreur de prédiction obtenue est faible.

Les hétéroéléments (soufre, azote et oxygène) et les métaux (nickel et vanadium) sont présents à de faibles teneurs. La spectroscopie infrarouge n'est généralement pas adaptée aux analyses de traces et ces éléments ne présentent pas de bandes spécifiques dans les spectres des produits lourds du pétrole. Ceci peut expliquer le peu de références pour le soufre et l'azote [54, 93] et l'absence de travaux pour l'oxygène, le nickel et le vanadium.

2.4 Bilan

Ce chapitre a permis de faire un bilan sur des travaux réalisés pour l'analyse rapide des produits lourds.

Nous avons pu déterminer que la spectroscopie PIR et la spectroscopie MIR en mode ATR, sont les techniques les plus utilisées. Cependant, il n'est pas possible de déterminer à partir de la littérature quelle méthode est la plus adaptée à notre application. Ainsi, il est envisagé de mener une étude de comparaison des spectroscopies MIR et PIR. Les travaux menés dans la littérature abordent l'analyse de tous types de produits lourds et de la plupart de leurs propriétés. Ces modèles de prédiction des propriétés des produits lourds ont majoritairement été développés à partir de l'algorithme PLS sur les spectres dérivés et en sélectionnant le domaine spectral en fonction de la propriété considérée. Ces modèles amènent à des performances satisfaisantes. La spectroscopie infrarouge est donc une technique de choix pour l'analyse rapide des produits lourds.

Les modèles présentés dans la littérature sont généralement développés sur un seul type de produits issus d'un procédé unique. Les gammes analytiques des propriétés considérées sont par conséquent le plus souvent assez restreintes. Or, l'objectif de notre travail est de répondre aux objectifs analytiques de l'ensemble des procédés de valorisation des produits lourds du pétrole, ce qui suppose une base d'échantillons plus diversifiée et d'étendue plus large.

Nous avons abordé dans ce chapitre que les propriétés des produits pétroliers sont très corrélées à la coupe pétrolière considérée (Partie 1.2.2). De plus, nous avons également vu que la qualité des effluents diffèrent selon les opérations de raffinage qu'ils ont subit (Partie 1.2.3). Ainsi, des échantillons de différentes coupes et provenant de plusieurs procédés de valorisation des produits lourds ont des compositions chimiques très différentes. Ces compositions chimiques diversifiées peuvent entraîner des signatures spectrales différentes. Ainsi, nous nous attendons à faire face à des problèmes lors du développement de l'analyse rapide des produits lourds que les méthodes classiques appliquées dans la littérature ne pourrait pas gérer. Ainsi, il est envisagé de faire appel à des méthodes moins communes telles que des procédures d'optimisation et des méthodes de combinaisons spectrales.

Tableau 2.1 – Résultats bibliographiques pour la détermination de la SARA

Références	Type d'échantillons	Technique spectroscopique	Algorithme	Étendue	Prétraitement	Domaine (cm^{-1})	A[a]	EC(V)[b]	EP[c]
Aske *et al.* [4]	18 bruts	ATR-IR	PLS	S : 24-83	Dérivée 1ère	4000-400	3	1,70	2,45
				A : 13-43			7	0,41	2,20
				R : 4-25			3	0,89	1,37
				@ : 0-12			7	0,33	1,29
		PIR	PLS	S : 24-83	Dérivée 1ère	9000-4500	8	1,13	2,78
				A : 13-43			6	0,85	2,39
				R : 4-25			7	0,54	1,41
				@ : 0-12			8	0,29	0,98
Blanco *et al.* [15]	66 bitumes	PIR	PLS	S : 3-10	Dérivée 2nd	6060-5560 4480-4020	5	0,3	0,6
				A : 34-55	Dérivée 2nd	6060-5560 4760-4020	5	1,6	2,1
				R : 18-34	Dérivée 2nd	6060-5560 4760-4020	5	1,6	2,2
				@ : 16-30	Dérivée 1ère	9000-4020	3	1	1,3
Hannisdal *et al.* [51]	20 bruts	PIR	PLS	S : 23-51	Dérivée 2nde	5900-5570	4	1,88	2,82
				A : 26-48	Dérivée 2nd	5900-5570 4670-4570	3	1,10	1,47
				R : 7-30	Dérivée 2nd	-	6	0,64	1,46
				@ : 0.2-13	Dérivée 1ère	9000-8680	2	0,38	0,44
		ATR-IR	PLS	S	Dérivée 1ère	1382-1470 2880-2960	3	1,29	1,84
				R	Dérivée 1ère	1550-1750	3	1,06	1,32
Hongfu *et al.* [54]	70 RA	ATR-IR	PLS	S : 18-68	Dérivée 2nd	1850-690	7	1,81	1,70
				A : 20-45	Dérivée 2nd	1850-690	9	1,36	1,13
				R : 6-39	Dérivée 2nd	1850-690	10	0,74	0,59
				@ : 20-45	Dérivée 1ère	1850-690	11	0,47	0,44
	68 RSV	ATR-IR	PLS	S : 14-51	Dérivée 2nd	1770-1150	8	1,73	1,78
				A : 32-43	Dérivée 2nd	1770-750	8	1,25	1,44
				R : 15-49	Dérivée 2nd	1770-750	11	0,59	0,81
				@ : 0.1-16	Dérivée 2ère	1770-750	12	0,44	0,47
	164 RH	ATR-IR	PLS	S : 32-66	Dérivée 2nd	1770-750	7	1,70	1,63
				A : 26-50	Dérivée 2nd	1770-750	8	1,35	1,41
				R : 7-22	Dérivée 2nd	1770-750	11	0,76	0,85
				@ : 0.1-25	Dérivée 1èr	1770-750	10	0,15	0,21
Satya *et al.* [93]	11 C_{12+}	PIR	PLS	S : 34-70	Dérivée	8300-4100	1	-	2,3
				A : 18-41	Dérivée		1	-	1,7
	11 C_{25+}			R : 6-27	Dérivée		1	-	1,1
				@ : 0-9.3	Dérivée		5	-	1,5

[a] : Nombre de composantes PLS
[b] : RMSEC ou RMSECV
[c] : RMSEP

Tableau 2.2 – Résultats bibliographiques pour la détermination de la viscosité

Références	Type d'échantillon	Technique Spectroscopique	Prétraitement	Domaine (cm^{-1})	Étendue (T°C)	Algorithme	A[a]	EC(V)[b]	EP[c]
Blanco *et al.* [15]	66 bitumes	PIR	Dérivée 1ère	9000-4000	474-11780 Po (60°C)	PLS	4	1297	924
						avec ln(visco)	13	165	334
					208-1056 cSt (135°C)	PLS	5	72	55
						avec ln(visco)	8	19	19
Blanco *et al.* [14]					158-1035 cSt (135°C)	ANN[d]			53
Hongfu06 *et al.* [54]	165 résidus hydrocraqués	ATR-IR	Dérivée 1ère	1710-670	11,6-57,0 $mm^2.s^{-1}$ (100°C)	PLS	8	1.34	1.75
Lima et Leite [70]	80 bitumes	PIR	Normalisation	11000-5800	890(moy : 2355) Po (60°C)	PLS	7	50	100
			+ Dérivée 1ère		226(moy : 207) cSt (135°C)		12	7.0	9,2

[a] : Nombre de composantes PLS
[b] : RMSEC ou RMSECV
[c] : RMSEP
[d] : Réseaux de Neurones Ariticiels

Tableau 2.3 – Résultats bibliographiques pour la détermination des autres propriétés

Propriétés	Références	Type d'échantillons	Technique Spectroscopique	Algorithme	Étendue	Prétraitement	Domaine (cm^{-1})	A[a]	EC(V)[b]	EP[c]
Densité	Blanco *et al.* [15] Blanco *et al.* [14]	66 bitumes	PIR	PLS ANN[d]	1,021 - 1,091	Dérivée 1ère	9000-4000	4 12	0,002 0,0020	0,002 0,0019
	Chung *et al.* [24]	81 RA	PIR	PLS	0,980 - 0,932	Dérivée 2nd	6060-4000	6	0,27	0,22
	Gilbert *et al.* [45]	111 DSV et RA	PIR	PLS	0,935 - 0,965	-	5000-4000		-	0,007
	Hidajat *et al.* [52]	105 bruts	PIR	PLS	0,7980 - 0,9038	MSC	10000-3700	10	-	0,0022
	Hongfu *et al.* [54]	124 résidus hydrocraqués	ATR-IR	PLS	0,9031 - 0,9692	Dérivée 1ère	1850-690	11	0,0006	0,0009
	Lima *et al.* [69]	88 huiles	PIR	PLS	0,814 - 0,927	Normalisation	11000-5800	7	0,004	0,0005
	Satya *et al.* [93]	11 C_{12+} 11 C_{25+}	PIR	PLS	0,866 - 0,999	Dérivée	8300-4100	7		0,028
API	Pasquini et Bueno [86]	79 bruts	PIR	PLS ANN[d]	23,7 - 30,4	Normalisation	5000-3900 6000-3700			0,24 0,44
Pt d'écoulement (°C)	Chung et Ku [25]	43 Huiles	PIR	PLS	-37 - -15	Dérivée 2nd	9090-6330	5	0,71	0,67
Le carbone	Gilbert *et al.* [45]	111 DSV et RA	PIR	PLS	5 - 6	-	5000-4000		0,24	
Conradson	Hongfu *et al.* [54]	165 résidus hydrocraqués	ATR-IR	PLS	3,52 - 10,77	Dérivée 1ère	1880-670	10	0,13	0,19
(%(w/w))	Satya *et al.* [93]	11 C_{12+} 11 C_{25+}	PIR	PLS	0,1 - 15,7	Dérivée	8300-4100	4		0.374
Asphaltènes C7	Blanco *et al.* [15]	66 bitumes	PIR	PLS	9,9 - 26,8	Dérivée 1ère	9000-4000	3	0,9	1,2
(%(w/w))	Long *et al.* [72]	sables bitumeux	PIR	PLS	0 - 20 20 - 100	-	10000-9000	2	0.20 1.1	0,23
Hydrogène (%(w/w))	Hongfu *et al.* [54]	69 résidus hydrocraqués	ATR-IR	PLS	11,19 - 12,79	Dérivée 1ère	1880-670	9	0.04	0,07
H/C	Satya *et al.* [93]	11 C_{12+} 11 C_{25+}	PIR	PLS	1,545 - 1,920	Dérivée	8300-4100	5		0,089
Carbone paraffinique (%(w/w)) Carbone naphténique (%(w/w)) Carbone aromatique (%(w/w))	Lima *et al.* [69]	88 huiles	PIR	PLS	42,1-94,7 4,2-94,7 0,4-15,1	Normalisation	11000-5800	5 6 4	1.2 0.5 0.5	1,4 1 0,5
soufre	Satya *et al.* [93]	11 C_{12+} 11 C_{25+}	PIR	PLS	0,05 - 2,35	Dérivée	8300-4100	2		0,035
(%(w/w))	Hongfu *et al.* [54]	65 résidus hydrocraqués	ATR-IR	PLS	0,28 - 2,2	Dérivée 1ère	1880-670	10	0.04	0,08
Azote	Satya *et al.* [93]	11 C_{12+} 11 C_{25+}	PIR	PLS	0 - 0,82	Dérivée	8300-4100	5		0.052
(%(w/w))	Hongfu *et al.* [54]	65 résidus hydrocraqués	ATR-IR	PLS	0,12 - 0,29	Dérivée 2nd	1880-670	7	0.01	0.02

[a] : Nombre de composantes PLS
[b] : RMSEC ou RMSECV
[c] : RMSEP
[d] : Réseaux de Neurones Ariticiels

3 Démarche analytique : matériels et méthodes associés

Le but de ce travail est de développer une analyse rapide des produits lourds du pétrole afin de proposer un suivi efficace de leurs procédés de valorisation. Les travaux réalisés dans la littérature montrent le potentiel des spectroscopies MIR et PIR pour l'analyse quantitative des différentes propriétés des produits lourds. Ce chapitre aura donc pour vocation de présenter, dans un premier temps, la démarche mise en œuvre pour le développement de modèles d'analyse multivariée. Nous décrirons dans les parties suivantes le matériel et les méthodes utilisées pour satisfaire cette démarche analytique. La base de données sera tout d'abord décrite. Nous exposerons ensuite le principe général des algorithmes génétiques ainsi que leur adaptation pour l'optimisation simultanée des prétraitements et de la sélection de variables. Les méthodes appliquées pour la fusion de données spectrales seront présentées par la suite. Enfin, les méthodes utilisées pour la comparaison de modèles seront détaillées.

3.1 La démarche analytique

La démarche de développement d'un modèle prédictif est illustrée sur la Figure 3.1. La première phase du développement d'une analyse multivariée consiste à élaborer une base de données. Le développement d'un étalonnage multivarié nécessite tout d'abord de collecter un lot d'échantillons dont les valeurs de référence des propriétés sont connues. Afin d'assurer la qualité des modèles, les échantillons d'étalonnage doivent être en nombre suffisant, représentatifs de la gamme analytique des propriétés et des échantillons qui

seront analysés dans le futur [57].

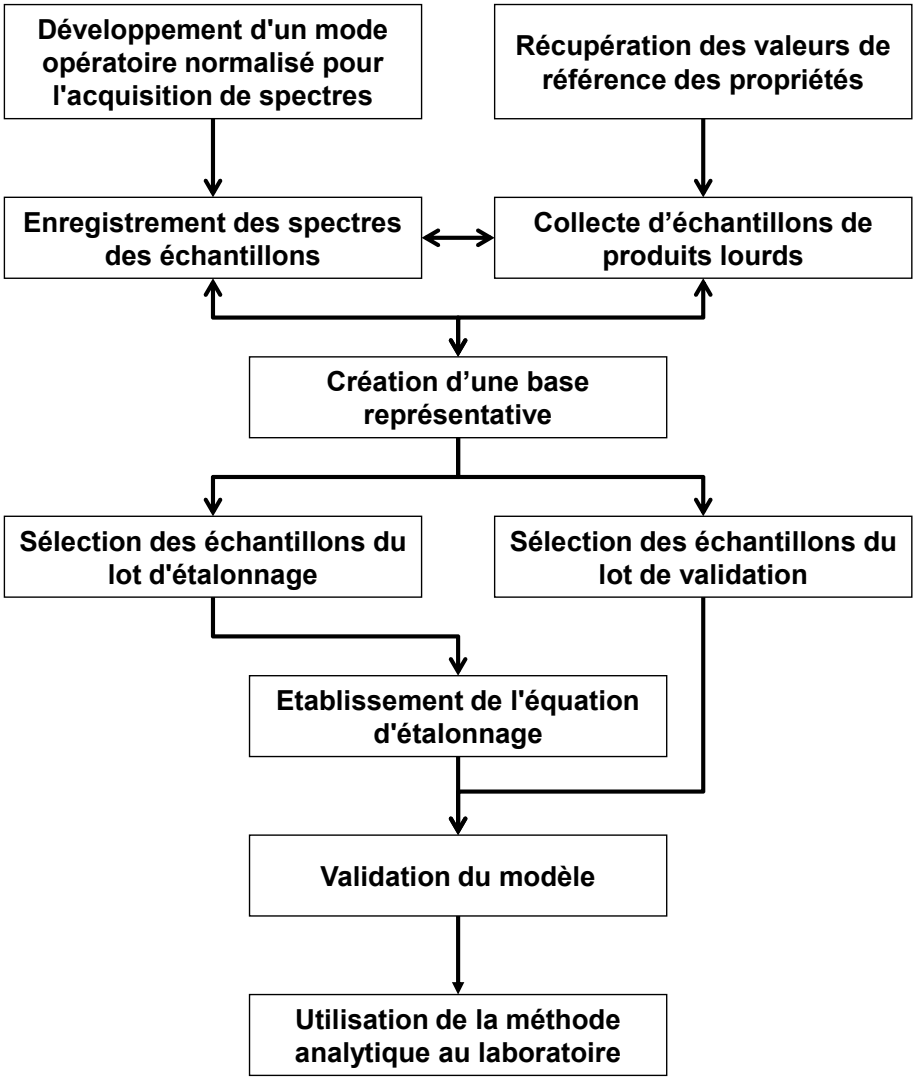


FIGURE 3.1 – Démarche de développement d'une analyse multivariée [12, 40]

La qualité d'un étalonnage multivarié dépend directement de la fidélité des mesures de référence des propriétés et de la mesure spectrale. En effet, les erreurs réalisées sur ces mesures seront introduites dans les modèles et vont dégrader leurs pouvoirs prédictifs [12]. Les méthodes d'analyses sont normalisées dans l'industrie pétrolière. Les techniques d'analyses sont donc bien établies et font l'objet de nombreux essais circulaires. Pour l'acquisition des spectres MIR et PIR, il est nécessaire d'optimiser les protocoles expérimentaux afin de maximiser la fidélité sur les mesures spectrales. Les modes opératoires

doivent ensuite être systématisés afin de s'assurer que tous les spectres sont mesurés dans les mêmes conditions et, de réduire ainsi les erreurs expérimentales. La première partie de ce chapitre sera donc consacrée à la présentation de la base d'échantillons et des modes opératoires pour l'acquisition des spectres MIR et PIR.

La deuxième phase du développement d'une analyse multivariée est l'établissement de l'équation d'étalonnage. L'un des points importants en amont de cette étape est la sélection des échantillons des lots d'étalonnage et de validation [27, 57]. Les échantillons du lot d'étalonnage vont servir au calcul de l'équation du modèle, tandis que ceux du lot de validation vont être utilisés pour l'évaluation de ses performances. Les échantillons du lot d'étalonnage doivent couvrir la gamme analytique de la base initiale et, notamment, les extrémités. Pour garantir une bonne évaluation des performances des modèles, les échantillons du lot de validation doivent couvrir le même intervalle. Ils doivent également se trouver dans la gamme définie par les échantillons d'étalonnage, afin d'éviter les problèmes d'extrapolation. Les méthodes appliquées pour la sélection des échantillons des lots d'étalonnage et de validation sont présentées en Annexe A.2.

Comme nous l'avons évoqué dans la Partie 2.2, les données spectrales brutes sont entachées de variations non désirées, c'est à dire, qui ne sont pas reliées aux variations de la propriété. Il en résulte que les données spectrales brutes ne sont pas forcément sous la forme la plus adaptée pour le calcul du modèle car, l'équation d'étalonnage est établie par une régression des valeurs de la propriété ($\mathbf{y}$) sur les absorbances du spectre ($\mathbf{X}$). Il est donc généralement nécessaire, d'une part, d'identifier les domaines spectraux d'intérêt et, d'autre part, d'appliquer des prétraitements mathématiques, qui visent à éliminer les interférences présentes dans les spectres [88].

A partir des connaissances spectroscopiques, il est possible de localiser les informations et d'identifier les perturbations dans les données spectrales. Néanmoins, il est difficile de définir *a priori* le domaine spectral et les méthodes de prétraitements qui amèneront au meilleur pouvoir prédictif. Il est donc nécessaire de procéder à une optimisation. L'optimisation consiste à déterminer le meilleur élément d'un ensemble, c'est à dire l'optimum, au sens d'un critère quantitatif donné. Dans le cadre du développement d'un modèle prédictif, cette optimisation est le plus souvent effectuée par une démarche d'essai-erreur.

Cette démarche consiste à tester un nombre fini et souvent restreint de prétraitements et de domaines spectraux. La solution adéquate est ensuite déterminée en fonction du pouvoir prédictif des modèles. Cependant, cette approche est longue, fastidieuse et aléatoire. De plus, il est difficile de savoir si l'optimum est atteint, même si le pouvoir prédictif du modèle est amélioré.

Ainsi, nous proposons d'évaluer le potentiel d'un algorithme d'optimisation. Les algorithmes d'optimisation sont basés sur le même principe que la démarche essai-erreur puisque leur but est de tester et d'évaluer des solutions potentielles. Cependant, l'intérêt des algorithmes d'optimisation est la possibilité, en un temps relativement court, de tester un nombre important de solutions potentielles. Ainsi, l'exploration des solutions potentielles est plus vaste ce qui permet une identification plus fidèle et plus rapide de l'optimum global du problème d'optimisation. Dans la Partie 3.3, nous décrirons une méthode pour l'optimisation simultanée des prétraitements et des variables à sélectionner par algorithmes génétiques [31].

Comme nous l'avons mentionné précédemment (Partie 2.1), les bandes d'absorption présentes en spectroscopies MIR et PIR ont la même origine car les bandes observées en spectroscopie PIR sont dues aux vibrations harmoniques et aux combinaisons des vibrations fondamentales. Néanmoins, l'instrumentation et l'interaction rayonnement-matière ne sont effectivement pas équivalents. Les variations spectrales ne sont donc pas identiques ce qui peut amener à des modèles de performances différentes (Partie 2.1.3). De plus, ces deux techniques comportent de l'information spécifique. Lors de ce travail, nous proposons donc, d'une part, de comparer les étalonnages multivariés développés à partir des spectroscopies MIR et PIR séparément. D'autre part, nous nous sommes intéressés aux méthodes permettant d'exploiter simultanément l'ensemble des informations spectrales à disposition. La Partie 3.4 aura donc pour vocation de décrire deux méthodes pour la fusion des données spectrales.

La troisième phase du développement d'une analyse multivariée est l'évaluation des performances du modèle prédictif. Les valeurs de propriétés des échantillons du lot de validation vont être prédites à partir de l'équation d'étalonnage. Différents critères statistiques vont ensuite permettre de déterminer les performances du modèle (Annexe A.5).

Pour chaque propriété, différentes approches ont été utilisées pour établir l'équation d'étalonnage. Il est donc nécessaire de développer une méthodologie pour les comparer et, ainsi, définir le modèle le plus performant. Afin d'appuyer nos conclusions, nous utiliserons donc un test statistique qui sera présentés dans la Partie 3.5. Une méthode de *bootstrap*, appliquée pour l'encadrement des valeurs prédites, sera également décrite.

La dernière phase du développement d'une analyse multivariée est l'utilisation en routine de l'analyse développée. Cette phase n'est pas développée ici car ce travail ne rentre pas dans le cadre de la thèse. Cependant, cette partie sera discutée dans les perspectives.

3.2 La base de données

Nous commencerons par présenter dans cette partie la base d'échantillons et nous discuterons de sa représentativité. Nous décrirons par la suite les spectromètres utilisés et les protocoles expérimentaux développés pour l'acquisition des spectres MIR et PIR.

3.2.1 Base d'échantillons

Le but des recherches menées sur les procédés de valorisation des produits lourds est de tester de nouveaux catalyseurs et d'optimiser les conditions opératoires des procédés. Ces recherches sont organisées en études qui peuvent durer de quelques semaines à plusieurs mois. Afin d'effectuer le suivi de ces études, des prélèvements d'effluents sont effectués quotidiennement. Pour des raisons de coût et de délais, le nombre d'analyses réalisées sur ces échantillons de suivi est restreint. A intervalles réguliers, des campagnes d'analyses plus complètes sont réalisées afin d'évaluer les performances des procédés.

Les caractéristiques (origine géographique, coupe pétrolière, résultats d'analyses) de tous les échantillons prélevés lors des études menées sont conservées. Ainsi, une base de données d'archive contenant les caractéristiques de plus de 6000 échantillons de produits lourds a été constituée. Cependant, ces échantillons ne sont pas tous conservés physiquement. Ainsi, le stock d'échantillons physiquement disponibles est restreint. De plus, pour les raisons abordées précédemment, toutes les propriétés ne sont pas disponibles pour tous les échantillons. Les échantillons issus de campagnes d'analyses plus complètes ont

donc plus d'intérêt dans le cadre de la constitution d'une base d'échantillons. En effet, le nombre de propriétés renseignées est plus élevé.

La méthodologie pour la constitution de la base est présentée sur la Figure 3.2. La base d'archive nous a tout d'abord permis de définir la diversité des produits analysés en termes d'origines géographiques, de procédés et de gammes analytiques des propriétés. Les échantillons d'intérêt ont ensuite été identifiés parmi le stock d'échantillons disponibles. Ils correspondaient, dans un premier temps, aux produits ayant fait l'objet d'un grand nombre d'analyses. En effet, cette démarche a permis de représenter suffisamment chaque propriété sans augmenter de manière trop significative le nombre d'échantillons à analyser. Une recherche plus spécifique a ensuite été réalisée pour couvrir au mieux les gammes analytiques des propriétés d'intérêt.

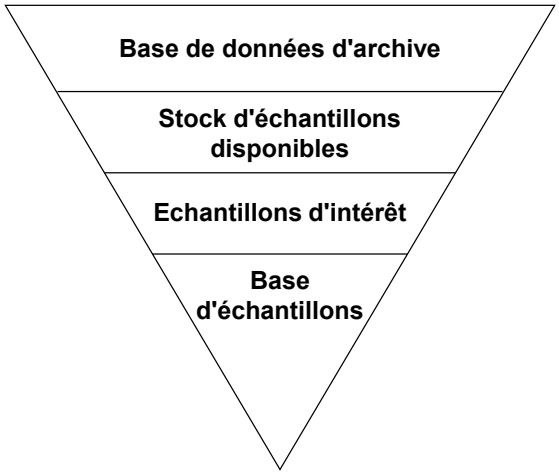


FIGURE 3.2 – Méthodologie pour la constitution de la base d'échantillons

La base ainsi sélectionnée comporte 230 échantillons (113 DSV et 117 RA). Ces échantillons sont tous des effluents. En effet, les produits bruts n'ont pas été pris en compte car les analyses de référence sont généralement préférées car ils font office de référence pour le suivi du procédé. Nous tenons à noter que les RSV ne sont pas présents dans cette base. Comme nous l'aborderons, l'acquisition des spectres nécessite de travailler en température. Le point d'ébullition initial des RSV (550°C) est très élevé par rapport aux DSV et aux RA (350°C). L'analyse des RSV a alors fait l'objet d'une autre étude qui

n'est pas décrite dans ce manuscrit.

Afin de déterminer si les échantillons sélectionnés sont représentatifs des produits lourds, nous comparerons la base d'échantillons à la base d'archive. Nous présenterons, tout d'abord, le nombre d'échantillons disponibles ainsi que la gamme analytique couverte par les échantillons de la base pour chaque propriété. Nous détaillerons par la suite la provenance des échantillons en termes d'origines géographiques et de procédés.

3.2.1.1 Propriétés

Le Tableau 3.1 synthétise le nombre d'échantillons disponibles et la gamme analytique couverte par les échantillons de la base pour chaque propriété. Les valeurs de certaines propriétés, telles que les teneurs en asphaltènes, peuvent être égales à zéro. Ainsi, le nombre d'échantillons qui ont des valeurs supérieures à zéro est indiqué entre parenthèses dans le tableau.

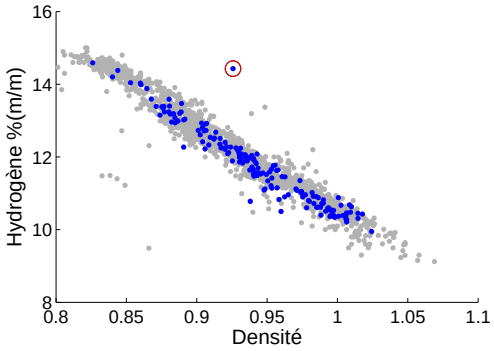
Pour la plupart des propriétés, plus d'une centaine d'échantillons sont disponibles. Comparé aux études menées dans la bibliographie (Partie 2.3), le nombre d'échantillons dans la base est assez important. Nous pouvons tout de même noter que la teneur en carbone et la teneur en carbones insaturés sont les deux propriétés les moins renseignées, avec 41 et 81 échantillons disponibles respectivement.

Tableau 3.1 – Caractéristiques de la base d'échantillons

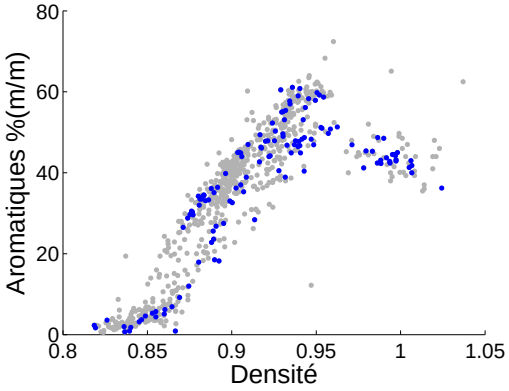
Propriété	**Nombre d'échantillons disponibles**	**Gamme analytique**
Densité	222	0,8186 - 1,0243
Indice de réfraction	117	1,4427 - 1,5482
Viscosité cinématique à 100°C (cSt)	153	3,45 - 793,19
Saturés %(m/m)	133	14,1 - 99,1
Aromatiques %(m/m)	133	0,7 - 61,1
Résines %(m/m)	133(132)[a]	0 - 34,5
Asphaltènes %(m/m)	133(24)[a]	0 - 12,8
Carbone Conradson %(m/m)	113	0,05 - 20,03
Asphaltènes C7 %(m/m)	171(91)[a]	0 - 14,2
Carbone %(m/m)	41	86,01 - 88,40
Hydrogène %(m/m)	171	9,95 - 14,59
Soufre %(m/m)	219	1.10^{-4} - 4,77
Azote (ppm)	189	0,1 - 11 800,0
Carbones insaturés %(m/m)	81	2,0 - 36,8
Nickel (ppm)	131(81)[a]	0 - 79
Vanadium (ppm)	135(85)[a]	0 - 328

[a] : Nombre d'échantillons dont la valeur de la propriété est supérieure à zéro

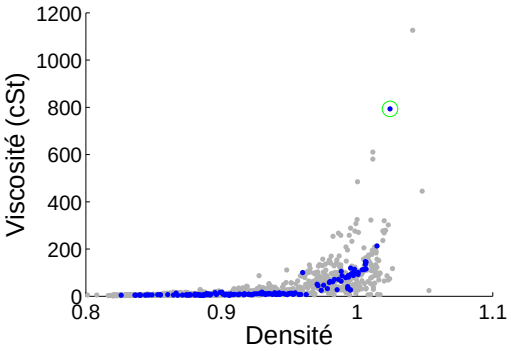
Afin de comparer la base d'échantillons à la base d'archive, le graphique des valeurs des propriétés en fonction de la densité a systématiquement été tracé. Sur la figure 3.3, nous illustrons la gamme analytique couverte par les échantillons de la base (en bleu) par rapport à ceux de la base d'archive (en gris) pour quatre propriétés : l'hydrogène, la fraction aromatique (SARA), la viscosité et les asphaltènes C7.



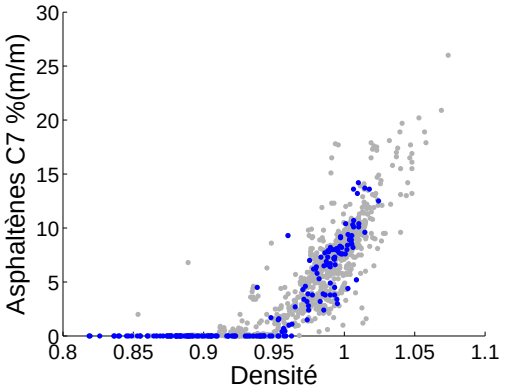
(a) Gamme analytique de la teneur en hydrogène



(b) Gamme analytique de la teneur en aromatiques (SARA)



(c) Gamme analytique de la teneur en viscosité



(d) Gamme analytique de la teneur en asphaltènes C7

FIGURE 3.3 – Gamme analytique de propriétés en fonction de la densité : Échantillons de la base (en bleu) - Échantillons atypiques (rouge) ou aux extrémités (vert) - Échantillons d'archive (en gris)

Nous pouvons observer que les échantillons de la base (en bleu) recouvrent bien la gamme analytique définie par les échantillons d'archive (en gris), hormis les extrémités de la gamme. Ces extrémités correspondent à des échantillons atypiques, qui ne sont plus

disponibles physiquement.

De plus, ces graphiques permettent d'identifier grossièrement les échantillons aberrants ou atypiques présents dans la base. En effet, nous pouvons, par exemple, remarquer un échantillon aberrant (entouré en rouge) sur le graphique de la teneur en hydrogène en fonction de la densité. Ce graphique illustre que la teneur en hydrogène et la densité sont corrélées. Nous pouvons alors supposer qu'un échantillon se trouvant en dehors de la droite définie par la majorité des échantillons d'archive est aberrant. Cet échantillon sera donc supprimé de la base avant le développement de l'équation d'étalonnage.

Enfin, ces graphiques permettent également d'identifier les valeurs de propriétés atypiques. Par exemple, nous pouvons observer, sur le graphique de la viscosité en fonction de la densité, que l'échantillon entouré en vert a une valeur de viscosité très élevée par rapport aux valeurs des autres échantillons de la base. De plus, très peu d'échantillons d'archive admettent des valeurs aussi élevées. Ainsi, la nécessité de l'introduction de cet échantillon pour le calcul de l'étalonnage multivarié peut se poser. Une décision sera prise en fonction de l'impact de la considération de cet échantillon sur les performances du modèle.

3.2.1.2 Origines géographiques

Les échantillons de la base sont des effluents de procédés de valorisation des produits lourds. Ils peuvent être obtenus à partir du raffinage de pétroles bruts, de produits ayant précédemment subi des opérations de raffinage ou de mélanges de produits lourds. Parmi les échantillons de la base, l'origine géographique de 172 échantillons a pu être déterminée. Ils ont été obtenus par le raffinage de 14 pétroles bruts différents. En revanche, 58 échantillons sont issus du raffinage de mélanges de produits provenant de plusieurs pétroles bruts. L'origine géographique de ces échantillons ne peut donc pas être déterminée. La Figure 3.4 illustre le nombre d'échantillons par origine géographique. Les pétroles provenant du Moyen Orient (Arabie Saoudite, Iran, Iraq) y sont largement représentés du fait de l'importance de la production dans cette région. Les sables bitumeux du Canada et les bruts provenant de Russie (Oural et Sibérie) sont également en nombre importants (25 échantillons chacun).

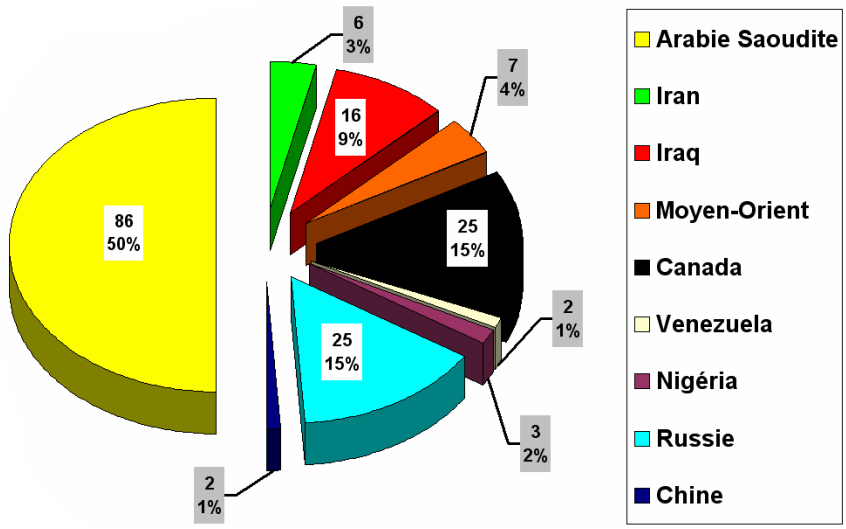


FIGURE 3.4 – Origine géographique des échantillons

3.2.1.3 Procédés

Les échantillons de la base sont des effluents provenant de trois procédés de valorisation des produits lourds : le procédé d'hydrocraquage, le procédé H-oil® et le procédé Hyvahl^{TM}. La Figure 3.5 illustre le nombre d'échantillons issus de ces procédés. Nous pouvons observer qu'environ 50% des échantillons sont issus du procédé H-oil®, 25% du procédé Hyvahl^{TM} et 25% du procédé d'hydrocraquage (HCK-DSV). Cette répartition est due au fait que le nombre d'échantillons provenant du procédé H-oil® est plus important dans le stock d'échantillons disponibles.

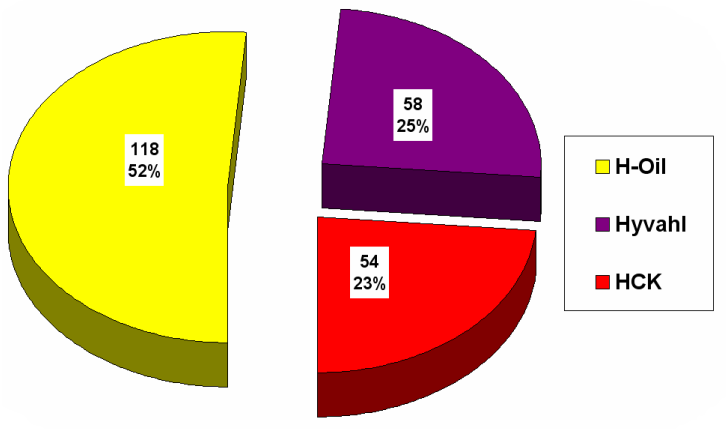


FIGURE 3.5 – Nombre d'échantillons provenant des différents procédés

3.2.2 Acquisition des spectres PIR

L'acquisition des spectres PIR des produits lourds est réalisée en mode transmission avec un trajet optique de 500 μm. Le spectromètre commercial ABB resid-IR utilisé a été spécialement conçu à cet effet (Figure 3.6). C'est un spectromètre à transformée de Fourier équipé d'un détecteur DTGS (*Deuterated Triglycine Sulfate*). Il permet de procéder à l'acquisition des spectres en transmission grâce à une cellule à circulation qui peut être maintenue en température, ce qui est nécessaire pour ces produits très visqueux. L'acquisition des spectres est réalisée à 100°C, température qui permet de s'assurer que les produits analysés sont liquéfiés et que les cristaux de paraffines sont solubilisés.

FIGURE 3.6 – Le spectromètre PIR

Le protocole expérimental consiste à introduire l'échantillon dans une coupelle en aluminium. Cette coupelle est ensuite placée dans un socle chauffant à une température de 100°C durant 10 minutes afin de liquéfier l'échantillon. La coupelle est, par la suite, re-

montée manuellement au niveau de la prise d'échantillon (une tige métallique plonge dans le produit). La circulation du produit dans la cellule puis vers le flacon de récupération se fait par aspiration à l'aide d'une pompe péristaltique. Après quelques secondes, le circuit est fermé en actionnant une vanne manuellement. L'échantillon est maintenu pendant 10 minutes dans la cellule thermostatée à 100 ± 0.1°C afin de procéder à sa stabilisation en température avant l'acquisition du spectre.

Le spectre est ensuite enregistré sur la gamme spectrale 12000 - 4000 cm^{-1}, avec une résolution d'environ 4 cm^{-1} et 100 scans sont réalisés. Le temps d'acquisition d'un spectre est de 4 minutes et 30 secondes. Deux spectres sont réalisés pour chaque échantillon sur deux prélèvements différents. La différence entre l'absorbance de ces deux spectres à chaque longueur d'onde est ensuite calculée. Si cette différence n'excède pas 5.10^{-3} unités d'absorbance sur la gamme 4500-4000 cm^{-1}, alors la mesure spectrale est jugée satisfaisante. Entre deux échantillons, la cellule est nettoyée par circulation de toluène et un spectre de référence est réalisé par basculement du faisceau sur un bloc de saphir. L'analyse d'un échantillon par spectroscopie PIR nécessite environ 35 minutes. Cette durée comprend l'acquisition du spectre de référence, la préparation de l'échantillon, l'acquisition des spectres sur les deux prélèvements et le nettoyage de la cellule.

3.2.3 Acquisition des spectres MIR

Les produits lourds sont analysés en spectroscopie MIR en mode ATR. Le spectromètre utilisé est un ThermoOptek Nicolet équipé d'un détecteur DTGS. L'accessoire ATR fonctionne en mono-réflexion. Il est muni d'un cristal en diamant entouré d'une plaque chauffante (Golden Gate d'Eurolabo).

Le protocole expérimental consiste à homogénéiser l'échantillon à température ambiante et à déposer un petit volume sur le cristal. La plaque chauffante permet l'acquisition du spectre à 100 ± 1 °C. Du fait du faible volume déposé, l'échantillon est instantanément stabilisé en température. Le spectre MIR est alors enregistré entre 4000 et 500 cm^{-1} avec une résolution de 4 cm^{-1} et 64 scans sont réalisés. Le cristal diamant coupe le signal à 650 cm^{-1}. Les spectres seront donc présentés par la suite sur le domaine 4000-650 cm^{-1}.

Pour chaque échantillon, deux spectres sont réalisés sur deux prélèvements. La mesure

spectrale est jugée satisfaisante si la différence entre eux, calculée longueur d'onde par longueur d'onde, ne dépasse pas 6.10^{-3} unités d'absorbance. Les spectres analysés sont très sensibles aux variations de la teneur en eau et en dioxyde de carbone dans l'atmosphère. Ainsi, un spectre de référence est réalisé avant chaque acquisition. Enfin, le cristal est nettoyé à l'acétone entre chaque acquisition. L'analyse d'un échantillon par spectroscopie MIR nécessite environ 10 minutes.

3.3 Les algorithmes génétiques (AG)

Nous avons abordé dans la Partie 3.1, la nécessité de sélectionner les domaines spectraux d'intérêt et d'appliquer des prétraitements afin de réduire les interférences physico-chimiques. De plus, Spiegelman *et al.* ont justifié théoriquement l'apport de la sélection de variables et ont montré qu'elle permettait d'améliorer les performances et la robustesse des modèles. De nombreuses méthodes de sélection de variables ont été développées. Elles ne sont pas décrites dans le détail ici mais le lecteur peut se référer à la "review" de Xiabo *et al.* [113]. Ces auteurs ont classé ces techniques de sélection de variables en cinq classes :

- les approches manuelles, basée sur les connaissances spectroscopiques (équivalent au choix des domaines spectraux d'intérêt) ;
- les méthodes de sélection univariées et séquentielles [97, 101] ;
- les méthodes "sophistiquées" : *successive projections algorithm* (SPA) [3] et *uninformative Variable Elimination* (UVE) [20] ;
- les algorithmes basés sur la sélection de variables par intervalle tels que l'*Interval Partial Least Squares* (iPLS) [83], la *windows PLS* [33, 59, 115] et l'*iterative PLS* [2, 22] ;
- les stratégies basées sur la recherche élaborée telles que le recuit simulé [58, 61, 98], les réseaux de neurones artificiels [16, 102] et les algorithmes génétiques (AG) [62, 64, 65].

Dans la littérature, ces méthodes de sélection de variables ont toutes montré leur efficacité pour l'amélioration du pouvoir prédictif d'étalonnage multivarié dans diverses applications. Lorsqu'une méthode de sélection de variables est appliquée, l'optimisation

du prétraitement et de la sélection de variables est généralement effectuée séparément. Nous avons abordé le fait qu'une démarche essai erreur peut-être longue et fastidieuse, en particulier lors de l'application d'une méthode de sélection de variables qui est chronophage. Un nombre restreint de prétraitements est alors testé. Cette démarche est justifiée car la méthode de prétraitement est généralement appliquée pour corriger des variations non-désirées, identifiables et souvent liées à des phénomènes physiques. La sélection de variables a ensuite pour objet d'identifier l'information spectrale qui est pertinente pour la description de la propriété considérée. Cependant, l'application d'un prétraitement peut, dans certains cas, également modifier l'information chimique. La sélection de variables est alors dépendante du choix du prétraitement appliqué en amont ce qui laisse penser qu'elle est potentiellement non-optimale.

Par conséquent, nous proposons d'évaluer le potentiel de l'application d'un algorithme pour l'optimisation simultanée du choix du (ou des) prétraitement(s) et de la sélection de variables. Il existe dans la littérature de nombreux algorithmes d'optimisation :

- les méthodes "Monte Carlo" [32, 48]
- les méthodes hybrides tels que la méthode des gradients [95]
- le recuit simulé [55, 61, 99]
- les algorithmes évolutionnistes tels que les AG [53]

Ces algorithmes d'optimisation n'ont cependant pas les mêmes performances en termes d'exploration (recherche globale) et d'exploitation (recherche locale) dans l'espace des solutions potentielles. En effet, les méthodes Monte Carlo permettent une bonne exploration puisque tout point a une probabilité identique d'être atteint. Cependant, aucune exploitation des résultats déjà obtenus n'est effectuée. Avec la méthode des gradients, l'exploration est moindre mais l'exploitation des données précédentes par l'intermédiaire des gradients permet une bonne recherche locale. Enfin, les algorithmes évolutionnaires offrent un bon compromis entre l'exploration et l'exploitation [7, 8].

Les AG appartiennent à la famille des algorithmes évolutionnistes (ou évolutionnaires). Ils sont inspirés de la théorie de l'évolution des espèces par la sélection naturelle. Selon la théorie de l'évolution par la sélection naturelle, au cours des générations, les êtres les plus adaptés à leur environnement tendent à survivre plus longtemps et à se reproduire plus

aisément. Ainsi, les caractéristiques génétiques conservés au sein d'une population donnée sont ceux qui sont les plus adaptés aux besoins de l'espèce vis à vis de son environnement. Ainsi, les espèces ont tendance à évoluer ("se perfectionner") au cours du temps car le patrimoine génétique des individus qui la composent "s'améliore". Les AG vont reproduire ce modèle d'évolution dans le but de trouver des solutions à un problème d'optimisation. Le principe des AG est donc de simuler l'évolution d'une population d'individus auxquels différents opérateurs génétiques sont appliqués et que l'on soumet, à chaque génération, à une sélection en fonction de leur adaptation à leur environnement. Les AG ont été introduits par Holland en 1975 [53]. Cependant, leur utilisation est récente du fait des ressources informatiques qu'ils requièrent. Les AG sont utilisés dans de nombreux domaines tels que le traitement d'image [104], l'optimisation d'emplois du temps, l'optimisation de design ou de contrôle de systèmes industriels [7], en chimie pour la modélisation moléculaire [30] et en chimiométrie pour la sélection de variables [62, 64, 65].

Nous décrirons dans cette partie le principe général de la procédure d'optimisation par AG. Nous présenterons par la suite l'application des algorithmes génétiques à l'optimisation simultanée des prétraitements et des variables à sélectionner.

3.3.1 Principe général des algorithmes génétiques

Les AG sont basés sur l'évolution des populations au cours des générations et sur les transmissions des caractéristiques par la génétique. Les AG sont itératifs. On peut distinguer deux phases dans la procédure d'optimisation des algorithmes génétiques : la phase d'initialisation et la phase d'évolution (Figure 3.7).

La phase d'initialisation comprend trois étapes : le codage du problème d'optimisation, la création de la population initiale et son évaluation.

Le codage du problème est une étape très importante qui peut conditionner la qualité de la procédure d'optimisation par AG. Il se présente sous une forme similaire au codage des caractéristiques des êtres vivants par la génétique. En effet, chaque solution au problème posé va être assimilée à un individu. Les caractéristiques de chaque solution sont codés sous forme d'une chaîne de bits qui représente le chromosome.

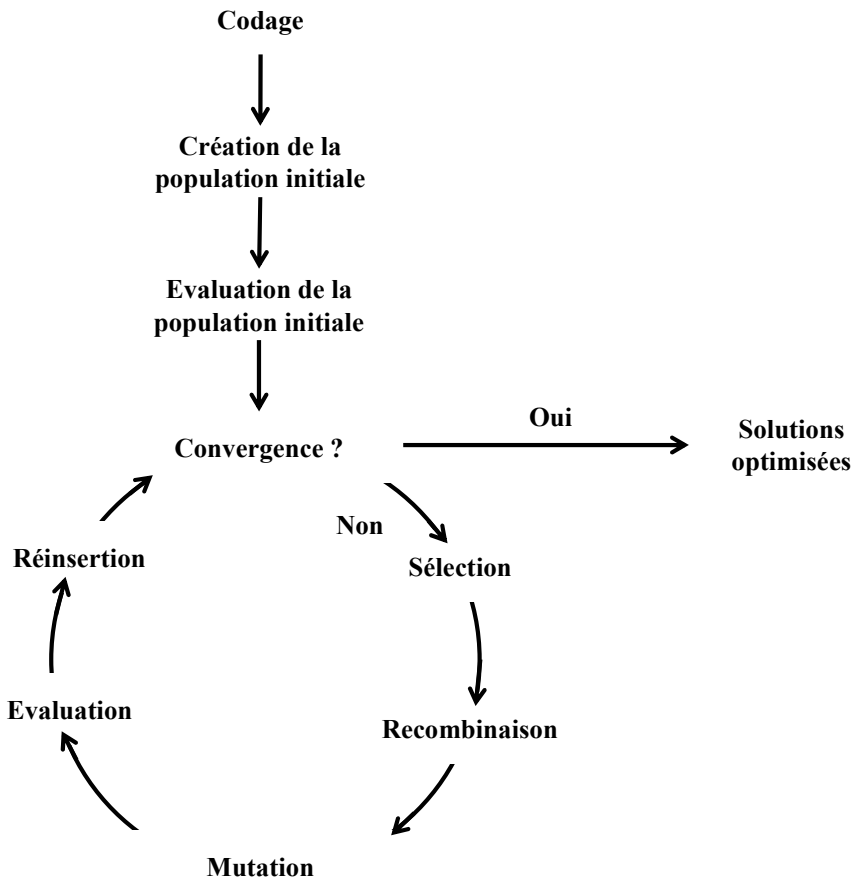


FIGURE 3.7 – Étapes de la procédure d'optimisation par algorithmes génétiques

Chaque bit du chromosome représente un gène et chaque gène ou groupe de gènes représente une variable du problème. Les allèles de chaque gène ou groupe de gènes correspondent à une valeur de la variable. Généralement, un codage binaire ou un codage à caractères multiples est utilisé.

La création de la population initiale nécessite de fixer plusieurs paramètres. Tout d'abord, le nombre d'individus qui composent la population doit être déterminé. La taille de la population est très important car, d'une part, pour un nombre d'individus trop petit, l'exploration sera faible. D'autre part, si le nombre d'individus est trop important, l'algorithme ne convergera pas en un temps raisonnable. La taille de la population peut-être assez différente en fonction des applications. Elle est généralement comprise entre 20 et 500 individus. Quand la taille de la population est définie, la première génération est créée par attribution d'une valeur aléatoire à chaque gène de chaque individu. La

population d'individus ainsi générée représente des solutions au problème d'optimisation. La phase d'évolution va ensuite avoir pour objectif d'optimiser la population des solutions afin de converger vers des solutions optimales.

L'évaluation consiste à estimer l'adaptation de chaque individu à son environnement. Pour les problèmes d'optimisation, l'adaptation à l'environnement correspond à une ou plusieurs fonction(s) à maximiser (ou minimiser).

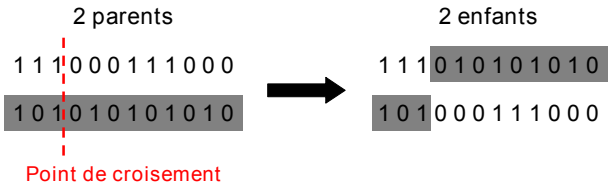
La phase d'évolution est itérative et comprend six étapes : la sélection, la recombinaison, la mutation, l'évaluation, la réinsertion et le contrôle de la convergence.

L'étape de sélection consiste à sélectionner les individus qui vont être utilisés pour créer la nouvelle génération. Par correspondance à la terminologie de la théorie de l'évolution, ils sont couramment appelés individus "parents". Il existe plusieurs façons de sélectionner ces individus :

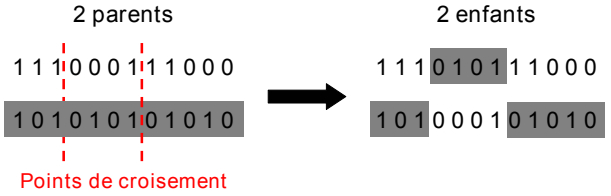
- la sélection aléatoire ;
- la sélection élitiste : les individus les plus adaptés sont sélectionnés ;
- La sélection par tournoi : les individus sont aléatoirement groupés par paire et le meilleur individu de chaque paire est sélectionné ;
- la sélection "roulette" : la probabilité de sélection d'un individu est proportionnelle à son adaptation [47].

Il est rare que la sélection aléatoire soit appliquée car la sélection de "parents" adaptés à leur environnement permet une convergence plus rapide et efficace vers des solutions optimales.

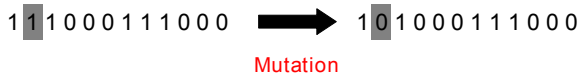
Lors de l'étape de recombinaison (ou de croisement) de nouveaux individus, communément appelés "enfants", sont créés par échanges d'une partie des chromosomes des "parents". Cette opération permet donc de créer une nouvelle génération avec pour objectif, conformément à la théorie de l'évolution, qu'elle soit globalement mieux adaptée à son environnement par rapport à la précédente. La recombinaison se réalise par paire d'individus qui sont choisis aléatoirement parmi les individus "parents". Le croisement entre chromosomes peut être simple (Figure 3.8a) ou multiple (Figure 3.8b). Le croisement simple consiste à fixer un point aléatoirement dans le chromosome et à échanger une des deux parties de la chaîne de gènes.



(a) Croisement simple entre deux chromosomes



(b) Croisements multiples entre deux chromosomes



(c) Mutation de l'allèle d'un gène

FIGURE 3.8 – Opération génétiques pour la création de nouveaux individus

Dans le cas de croisements multiples, il y a plusieurs points de croisements et plusieurs parties du chromosome sont échangées. Un taux de croisement, qui correspond au pourcentage de nouveaux individus créés par génération, est défini.

L'étape de mutation consiste à changer aléatoirement la valeur de certains gènes (Figure 3.8c). Un pourcentage de mutation par génération est défini et il est généralement compris entre 0,1 et 1%. Cette opération a plusieurs objectifs : éviter la convergence trop rapide de l'algorithme ce qui permet d'éviter les minima locaux et d'explorer des solutions proches des solutions optimisées sans altérer de manière trop significative leur adaptation à l'environnement.

L'adaptation des individus ainsi créés après recombinaisons et mutations est ensuite évaluée de la même manière que lors de la phase d'initialisation.

L'étape de réinsertion a pour but de choisir les individus qui vont créer la nouvelle génération. Les méthodes de réinsertion qui peuvent être utilisées sont les mêmes que celles appliquées pour l'étape de sélection.

Les différentes étapes de la phase d'évolution sont ainsi répétées jusqu'à convergence de

l'algorithme. Le contrôle de la convergence consiste à examiner le pourcentage d'individus dupliqués présents dans la nouvelle génération. Si le pourcentage maximum autorisé est atteint, alors l'algorithme est stoppé. Les individus de la dernière génération sont les solutions optimisées du problème. Un nombre maximum de générations est également défini qui sert également de critère d'arrêt de l'algorithme.

3.3.2 Avantages et limitations des AG

Les principaux avantages des AG sont, tout d'abord, qu'ils sont adaptables à de nombreux problèmes. De plus, comme nous l'avons abordé, les AG offrent un très bon compromis entre exploration et exploitation. Enfin, les AG convergent vers une population de solutions optimales ou proches de l'optimum. Ainsi, les AG offrent la possibilité de mettre en œuvre une démarche ou des critères pour la définition de la solution finale. Néanmoins, ils présentent également des limitations. En effet, la procédure des AG sollicite la définition de nombreux paramètres (taille de population, probabilité de croisement et de mutation, méthode de sélection des individus pour l'étape de sélection et de réinsertion ...) qui peuvent être délicats à régler. De plus, les AG requièrent le calcul de la fonction d'adaptation un très grand nombre de fois. Ce nombre de calculs peut s'avérer problématique lorsque le coût de calcul (ressources systèmes ou temporelles) de la fonction d'adaptation est important. Enfin, il est nécessaire de renouveler la procédure d'optimisation plusieurs fois afin d'évaluer l'influence de l'initialisation et de s'assurer de la robustesse de la convergence.

3.3.3 Application des AG à la co-optimisation

Afin d'adapter les AG à un problème d'optimisation, deux étapes sont à définir : le codage du problème pour qu'il puisse être résolu par AG et la fonction d'évaluation de l'adaptation des individus. Les AG ont été introduits en chimiométrie par Leardi *et al.* [66] pour la sélection de variables dans le cadre de développement d'étalonnages multivariés. Depuis, de nombreux travaux d'optimisation ont été réalisés en chimiométrie à partir des AG [35, 34, 62, 65, 67, 92]. Nous présenterons, dans cette partie, une procédure

d'optimisation simultanée (ou co-optimisation) du prétraitement et de la sélection de variables [31] par AG.

3.3.3.1 Le codage de la co-optimisation

Dans le cadre de l'optimisation de la sélection de variables, le codage est réalisé en binaire [66]. Le chromosome de chaque individu est une chaîne de bits où chaque bit représente une variable du spectre. Le bit est assigné à 1 lorsque la variable est sélectionnée et à 0 lorsqu'elle ne l'est pas (Figure 3.9).

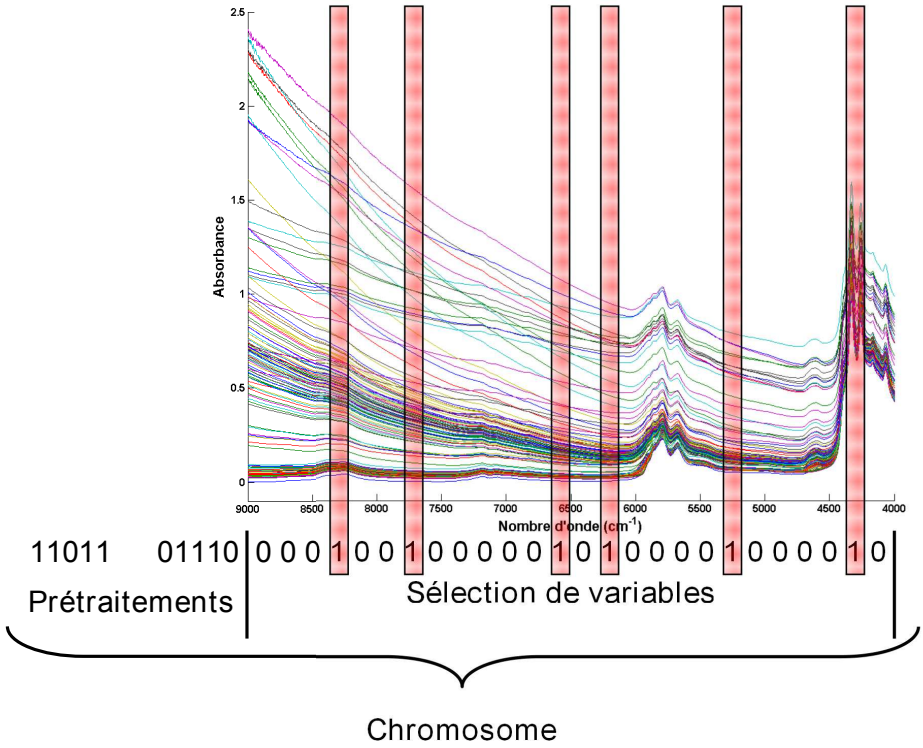


FIGURE 3.9 – Représentation du codage des prétraitements et de la sélection de variables pour la co-optimisation par AG ; Le chromosome représenté ici correspond à l'application de deux prétraitements : un centrage par la médiane (11011), d'une dérivée 1ère calculée sur une fenêtre de 15 variables et à partir d'un polynôme d'ordre 2 (01110) (Tableau 3.2). De plus, ce chromosome correspond à la sélection de six zones spectrales (en rouge) contenant chacune plusieurs variables (la sélection de variables est représentée sur les spectres bruts)

Un paramètre est introduit lors de la création de la population initiale pour fixer le pourcentage de variables qui seront sélectionnées pour chaque individu. En effet, si le taux de variables sélectionnées est trop petit, l'exploration de l'espace des solutions est faible. Au contraire, s'il est trop grand, la procédure d'optimisation est longue et l'algorithme ne converge pas. Nous pouvons noter que ce paramètre est étroitement lié à la taille de la population car il influe de la même manière sur l'exploration de l'espace de données et sur la convergence de l'algorithme.

Afin d'adapter le problème à la co-optimisation, l'information concernant l'application des prétraitements est codée dans la première partie du chromosome (Figure 3.9). Pour cela, 32 prétraitements ont été codés sur 5 bits ($2^5 = 32$). Ces prétraitements ont été sélectionnés afin de couvrir les méthodes les plus couramment appliquées en chimiométrie (Tableau 3.2) : les méthodes de filtrage, de correction de ligne de base, de normalisation et de dérivation. Le principe et l'intérêt de ces prétraitements sont présentés en Annexe A.1. Les différentes méthodes de centrage et de mise à l'échelle sont également implémentées. L'algorithme permet de combiner jusqu'à 4 prétraitements qui sont appliqués aux spectres dans le même ordre qu'ils apparaissent dans le chromosome (Figure 3.9). Afin de permettre à l'utilisateur de fixer le nombre maximum de prétraitement pouvant être appliqués durant l'optimisation, un paramètre p a été introduit ($p = 1$ à 4). La méthode développée permet également l'optimisation du prétraitement en utilisant le spectre entier ou un domaine spectral continu. Pour cela, les chromosomes sont uniquement constitués des gènes correspondant au codage des prétraitements.

3.3.3.2 La fonction d'évaluation

Dans le cadre du développement de modèle de prédiction, la fonction d'adaptation des individus est l'erreur de prédiction. Afin d'évaluer cette erreur de prédiction, pour chaque individu, le(s) prétraitement(s) et la sélection de variables sont appliqués aux spectres. Un modèle PLS est ensuite développé. Le nombre optimal A de facteurs PLS est déterminé par validation croisée (Annexe A.4).

Tableau 3.2 – Les 32 prétraitements disponibles lors de la co-optimisation par algorithmes génétiques (**K** : ordre du polynôme ; **F** : Nombre de points dans le filtre ; **M** : ordre de la dérivée

Catégorie	**Méthodes**	**Paramètres**			**Codage**
		K	F	M	
	Pas de prétraitement				00000
Filtrage	Savitzky-Golay smoothing (K, F)	2	15		00001
		3	15		00010
		2	7		00011
		3	7		00100
		2	21		00101
		3	21		00110
Normalisation	Multiplicative Scatter Correction (MSC)				01010
	Normalisation de la somme à l'unité				01011
	Standard Normal Variate (SNV)				01100
Dérivation	Savitzky-Golay derivative (K, F, M)	2	7	1	01101
		2	15	1	01110
		2	21	1	01111
		3	7	1	10000
		3	15	1	10001
		3	21	1	10010
		2	7	2	10011
		2	15	2	10100
		2	21	2	10101
		3	7	2	10110
		3	15	2	10111
		3	21	2	11000
Correction de la ligne de base	Detrend (K)	1			00111
		2			01000
		3			01001
	Weighted Least Square Baseline (K)	1			11101
		2			11110
		3			11111
Centrage et mise à l'échelle	Autoscale				11001
	Centrage par la moyenne				11010
	Centrage par la médiane				11011
	Square root mean scale				11100

Ce choix est effectué automatiquement lors de la procédure d'optimisation. En effet, le nombre de facteurs est fixé à A si le gain en RMSECV est inférieur à un pourcentage déterminé entre les modèles à A facteurs et A+1 facteurs (5% par exemple). La valeur d'adaptation de chaque individu correspond à la RMSECV obtenue pour A facteurs PLS.

Lors de cette étape d'évaluation, un des points importants, pour l'estimation de l'erreur de prédiction de chaque individu, est le choix de la méthode de validation croisée à appliquer. Les différentes méthodes sont présentées dans en Annexe A.4. Nous discuterons

du choix de la méthode lors de la présentation des paramètres utilisés pour la procédure d'optimisation par AG (Partie 4.2).

3.4 La fusion de données spectrales

Différentes méthodes pour l'exploitation simultanée des spectres MIR et PIR sont présentées dans cette partie. La manière la plus intuitive pour développer un étalonnage multivariée à partir de plusieurs bases spectrales est de les concaténer dans une matrice unique $\mathbf{X}$ et de développer un modèle PLS. Cependant, l'interprétation des résultats à partir de cette méthode est difficile car il n'est pas possible de définir la contribution exacte de chaque bloc dans la description de la réponse $\mathbf{y}$. Ainsi, des méthodes de fusion, appelées méthodes *multiblock*, ont été introduites dans la littérature. La philosophie de ces méthodes est de séparer les variables en différents blocs afin d'améliorer l'interprétation des étalonnages multivariés [108].

Dans cette partie, nous présenterons deux méthodes *multiblock* qui seront appliquées pour l'exploitation simultanée des spectres MIR et PIR : la méthode *multiblock* PLS (MB-PLS), qui traite les blocs en parallèle [108] et la *serial* PLS (S-PLS), qui les traite en série [11]. Ces deux méthodes sont des méthodes prédictives. Pour chaque méthode, la procédure mathématique sera illustrée en faisant intervenir deux blocs de variables, les spectres MIR et PIR, notés $\mathbf{X}_b$ et une réponse unique $\mathbf{y}$. Dans le cadre de la S-PLS, la différenciation entre les blocs est nécessaire, ils seront alors notés $\mathbf{X}_1$ et $\mathbf{X}_2$.

Dans la littérature, la MB-PLS a été appliquée pour combiner des spectres MIR et PIR pour l'analyse de graines de soja [18]. Elle a également été utilisée dans le cadre de suivi de procédé [73, 63, 74]. La S-PLS a été appliquée pour combiner les spectres MIR et PIR pour l'analyse de graines de soja [18] et de produits pétroliers (essences et gasoils) [41].

3.4.1 *Multiblock* PLS

Le concept de séparer les variables en différents blocs a été introduit par WOLD *et al.* [110]. Il existe de nombreuses méthodes *multiblock*. Le principe de ces méthodes consiste à combiner les différents blocs en concaténant les scores ACP ou PLS, calculés sur chaque bloc $\mathbf{X}_b$, dans une matrice unique $\mathbf{T}$. Cette matrice $\mathbf{T}$, qui regroupe les informations des différents blocs, est ensuite utilisée pour calculer une ACP ou une PLS, selon les cas. Deux niveaux sont alors distingués dans les méthodes MB-PLS : le *sub-level* et le *super-level* qui correspondent respectivement aux calculs effectués sur les différents blocs $\mathbf{X}_b$ et sur la matrice $\mathbf{T}$. Les méthodes *concensus PCA* (CPCA) [110] et *hierarchical PCA* (HPCA) [111] sont dérivées de l'ACP. Elles ont donc pour but de visualiser les données. Les méthodes HPLS [111] et *multiblock* PLS (MB-PLS) [105, 107] sont, quant à elles, issues d'une adaptation de la PLS. Elles ont donc pour objectif de développer un étalonnage multivarié. Ces méthodes ne sont pas toutes présentées ici mais un historique des différentes implémentations est disponible dans la publication de WESTERHUIS *et al.* [108]. L'algorithme de la méthode *Multiblock* PLS décrite dans cette partie est celui développé dans la publication de WESTERHUIS *et al.* [107]. Le principe de la MB-PLS est résumé sur la Figure 3.10 et l'algorithme est exposé ci dessous.

Pour a = 1 à A

$$\mathbf{w}_b = \frac{\mathbf{X}_b^T \cdot \mathbf{y}}{\mathbf{y}^T \cdot \mathbf{y}} \quad \% \text{ Calcul des poids des variables de chaque bloc} \tag{3.1}$$

$$\|\mathbf{w}_b\| = 1 \quad \% \text{ Normalisation de } \mathbf{w}_b \tag{3.2}$$

$$\mathbf{t}_b = \mathbf{X}_b \cdot \mathbf{w}_b \quad \% \text{ Calcul des scores de chaque bloc} \tag{3.3}$$

$$\mathbf{T} = [\mathbf{t}_1 \dots \mathbf{t}_2] \quad \% \text{ Combinaison des scores des blocs} \tag{3.4}$$

$$\mathbf{w_T} = \frac{\mathbf{T}^T \cdot \mathbf{y}}{\mathbf{y}^T \cdot \mathbf{y}} \quad \% \text{ Calcul des super-poids} \tag{3.5}$$

$$\|\mathbf{w_T}\| = 1 \quad \% \text{ Normalisation de } \mathbf{w_T} \tag{3.6}$$

$$\mathbf{t_T} = \mathbf{T} \cdot \mathbf{w_T} \quad \% \text{ Calcul des super-scores} \tag{3.7}$$

$$\mathbf{q} = \frac{\mathbf{y}^T \cdot \mathbf{t_T}}{\mathbf{t_T}^T \cdot \mathbf{t_T}} \quad \% \text{ Calcul des poids de } \mathbf{y} \tag{3.8}$$

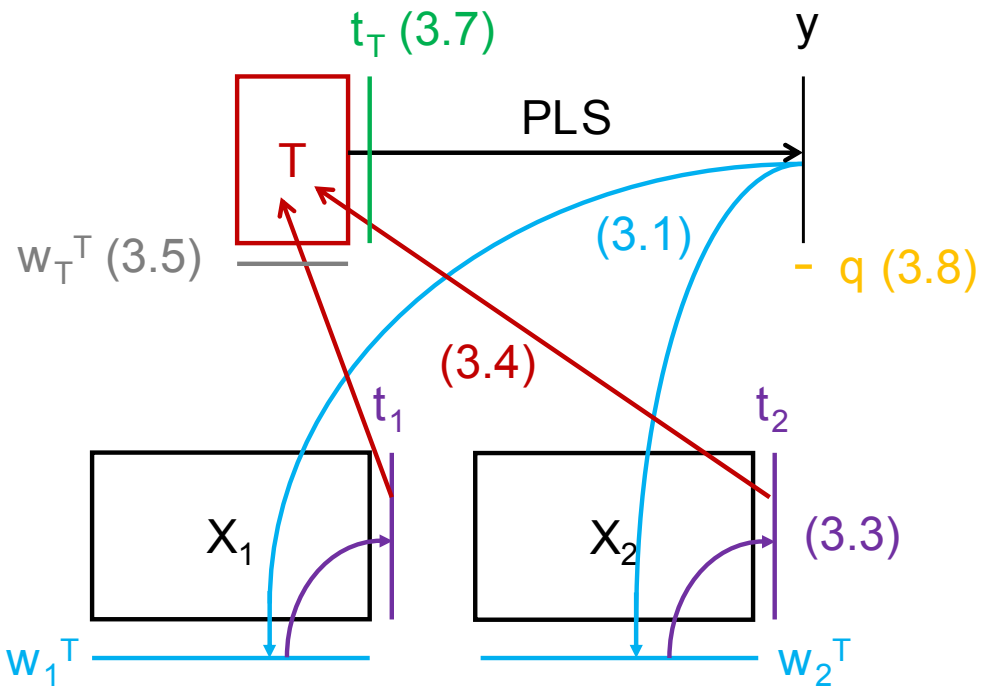


FIGURE 3.10 – Principe de la MB-PLS ; les étapes de l'algorithme sont indiquées sur la figure et correspondent à une couleur (Equations 3.1 à 3.8). Les étapes 3.2 et 3.6, qui correspondent à la normalisation de $\mathbf{w}_b$ et $\mathbf{w}_T$, ne sont pas indiquées sur cette figure

La MB-PLS consiste à calculer un premier cycle PLS entre chaque bloc $\mathbf{X}_b$ et la réponse $\mathbf{y}$ (Equations 3.1 et 3.2). Ce que nous dénommons ici, cycle PLS, correspond au calcul effectué pour chaque dimension dans la régression PLS (Annexe A.3). Cette première étape permet de définir les scores PLS $\mathbf{t}_b$ de chaque bloc. Il faut noter que la norme des poids des variables de chaque bloc $\mathbf{w}_b$ est ramenée à l'unité, afin de donner le même poids à chaque bloc. Les scores $\mathbf{t}_b$ sont ensuite concaténer dans une matrice $\mathbf{T}$. Le modèle prédictif est ensuite calculé en réalisant un deuxième cycle PLS entre cette matrice des scores $\mathbf{T}$ et la réponse $\mathbf{y}$ (Equations 3.5 à 3.8). Ces deux cycles PLS sont répétés pour chaque facteur a de la MB-PLS jusqu'à ce que le nombre de facteurs optimal A soit atteint.

De manière équivalente au cycle de la PLS, après chaque dimension, la partie expliquée par le facteur a est retirée aux matrices de chaque bloc $\mathbf{X}_b$ et à la réponse $\mathbf{y}$. Cette opération est appelée *deflation* en anglais. Nous utiliserons donc ce terme ici. La *deflation* permet d'assurer que l'information décrite par une dimension n'apparaisse pas dans les dimensions suivantes. Cette étape assure donc l'orthogonalité des dimensions et est à

l'origine de capacité de la PLS à gérer les cas où les variables sont très corrélées. En PLS, la *deflation* sur $\mathbf{X}$ et $\mathbf{y}$ est calculée de la façon suivante (Partie A.3) :

$$\mathbf{X} = \mathbf{X} - \mathbf{t} \cdot \mathbf{p}^T \tag{3.9}$$

$$\mathbf{Y} = \mathbf{y} - \mathbf{t} \cdot \mathbf{q}^T \tag{3.10}$$

$$\tag{3.11}$$

avec $\mathbf{t}$: matrice des scores PLS

$\mathbf{p}$: matrice des loadings des variables $\mathbf{X}$

$\mathbf{q}$: matrice des loadings de $\mathbf{y}$

En MB-PLS, deux cycles PLS sont réalisés. Deux manières de procéder à la *deflation* sont donc possibles : en utilisant les scores $\mathbf{t}_b$ du premier cycle ou les scores $\mathbf{t_T}$ du second cycle. Dans la version de la MB-PLS développée par Wangen et Kowalski [105], la *deflation* est effectuée en utilisant les scores $\mathbf{t}_b$ de chaque bloc. Par conséquent, les dimensions des premiers cycles PLS sont orthogonales. Le problème de cette méthode de *deflation* est que la totalité de la direction de $\mathbf{t}_b$ est soustraite à $\mathbf{X}_b$. Or, d'après l'équation 3.5, seule la partie $\mathbf{w_T}(\mathrm{b}) \cdot \mathbf{t}_b$ est utilisée dans le deuxième cycle PLS. Ici, $\mathbf{w_T}(\mathrm{b})$ représente les super-poids correspondant au bloc b. Cette approche peut donc amener à une perte d'information [109]. De plus, les scores $\mathbf{t_T}$ peuvent être corrélés. En effet, l'orthogonalité des dimensions du deuxième cycle PLS n'est donc pas assurée. Or, étant donné que les coefficients PLS du modèle prédictif sont calculés à partir du deuxième cycle PLS, réaliser la *deflation* à partir des scores $\mathbf{t}_b$ peut amener à des prédictions non-optimales [109]. La méthode décrite par Westerhuis *et al.* [108] procède donc à la *deflation* à partir des super-scores $\mathbf{t_T}$ de la manière suivante :

$$\mathbf{p_{T}}_b = \frac{\mathbf{X}_b^T \cdot \mathbf{t_T}}{\mathbf{t_T}^T \cdot \mathbf{t_T}} \tag{3.12}$$

$$\mathbf{X}_b = \mathbf{X}_b - \mathbf{t_T}^T \cdot \mathbf{p_{T}}_b^T \tag{3.13}$$

$$\mathbf{y} = \mathbf{y} - \mathbf{t_T} \cdot \mathbf{q}^T \tag{3.14}$$

Le principal avantage de cette méthode est que le calcul en deux cycles permet ensuite de déterminer la contribution de chaque bloc pour la description de la réponse $\mathbf{y}$. De plus, l'interprétation de la contribution des données spectrales combinées est également possible. Par conséquent, l'interprétation du modèle est plus complète. Enfin, il a été montré que le pouvoir prédictif du modèle n'est pas altéré par la présence de deux cycles [109].

3.4.2 *Serial PLS*

La méthode *Serial* PLS (S-PLS) a été introduite par Berglund et Wold [11]. Le principe de la S-PLS est le suivant : "Les modèles de chaque bloc sont calculés à partir des résidus de $\mathbf{y}$ du modèle précédent", chaque modèle étant développé en utilisant une régression PLS. La philosophie de cette méthode diffère donc de la MB-PLS car les blocs sont traités en série. De plus, contrairement aux méthodes *multiblock* pour lesquelles les données sont combinées dans une matrice contenant les scores, en S-PLS les données sont uniquement reliées par la réponse $\mathbf{y}$. La régression S-PLS s'écrit :

$$\mathbf{X}_1 = \mathbf{T}_1 \cdot \mathbf{P}_1^T + \mathbf{E}_1 \tag{3.15}$$

$$\mathbf{X}_2 = \mathbf{T}_2 \cdot \mathbf{P}_2^T + \mathbf{E}_2 \tag{3.16}$$

$$\mathbf{y} = \mathbf{T}_1 \cdot \mathbf{q}_1^T + \mathbf{T}_2 \cdot \mathbf{q}_2^T + \mathbf{f} \tag{3.17}$$

avec $\mathbf{T}$: matrice des scores PLS

$\mathbf{P}$: matrice des loadings des variables

$\mathbf{q}$: matrice des loadings de la réponse

$\mathbf{E}$: Residus sur $\mathbf{X}$

$\mathbf{f}$: Résidus sur $\mathbf{y}$

De manière équivalente, le modèle de régression S-PLS peut également s'écrire selon l'Equation 3.18 en faisant intervenir les coefficients de régression β.

$$\mathbf{y} = \mathbf{X}_1 \cdot \beta_1 + \mathbf{X}_2 \cdot \beta_2 + \mathbf{f} \tag{3.18}$$

L'algorithme de la S-PLS est un algorithme itératif qui peut se diviser en plusieurs étapes. Nous illustrons cet algorithme dans le cas où deux blocs de variables, notés $\mathbf{X}_1$ et $\mathbf{X}_2$, sont considérés (Figure 3.11) :

1. Initialisation : $\mathbf{f}_2 = \mathbf{y}$
2. Calculer le 1er modèle PLS avec $\mathbf{X}_1$ et $\mathbf{f}_2$
3. Calculer les résidus du 1er modèle : $\mathbf{f}_1 = \mathbf{y}$ - $(\mathbf{T}_1 \cdot \mathbf{q}_1^T)$
4. Calculer le 2nd modèle PLS avec $\mathbf{X}_2$ et $\mathbf{f}_1$
5. Calculer les résidus du 2nd modèle : $\mathbf{f}_2 = \mathbf{y}$ - $(\mathbf{T}_2 \cdot \mathbf{q}_2^T)$
6. Répéter les étapes 2 à 5 jusqu'à la convergence

La convergence de l'algorithme est atteinte lorsque la somme des différences quadratiques entre les résidus $\mathbf{f}_2$ de deux itérations successives est inférieure à un seuil.

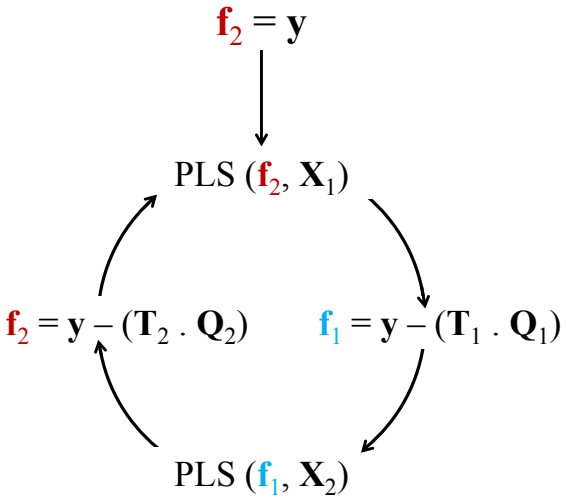


FIGURE 3.11 – Principe de la S-PLS

Le principal avantage de cette méthode est que la variance de $\mathbf{y}$ qui n'est pas expliquée par un bloc peut potentiellement l'être par un autre bloc. On peut donc s'attendre à ce que la variance totale de $\mathbf{y}$ expliquée par cette méthode soit plus grande qu'avec d'autres méthodes telles que la PLS ou la MB-PLS. De plus, la S-PLS étant composée de régressions PLS, les outils d'interprétation sont identiques à ceux de la PLS (pourcentage de variance expliquée, coefficients de régression β, poids des variables $\mathbf{P}$, scores $\mathbf{T}$). Enfin, ces outils

d'interprétation étant disponibles pour chaque bloc, les résultats obtenus permettent de déterminer la contribution de chaque bloc pour la description de la réponse $\mathbf{y}$.

La méthode S-PLS est donc intuitive et facile à mettre en œuvre. Bien que la phase itérative puisse paraitre superflue, BERGLUND et WOLD ont montré qu'elle permettait d'obtenir de meilleurs résultats [11]. Ils interprètent cette observation par le fait que, lorsque l'itération est effectuée, tous les modèles sont calculés à partir de résidus de $\mathbf{y}$. Dans le cas contraire, le premier bloc est calculé sur les données brutes de $\mathbf{y}$, ce qui peut influencer les résultats.

La méthode S-PLS permet d'affecter un nombre de facteurs différent pour chaque bloc. BERGLUND et WOLD précisent néanmoins qu'ils doivent être fixés indépendamment. Ils proposent pour cela d'utiliser la validation croisée et de tester toutes les combinaisons possibles. Si un modèle S-PLS est calculé en considérant deux blocs, les résultats correspondent à une matrice carrée contenant toutes les valeurs d'erreur de prédiction en validation croisée où les lignes correspondent aux nombres de composantes pour le premier bloc et les colonnes à celles du deuxième bloc.

Nous pouvons noter que l'ordre des blocs est également important car l'initialisation est réalisée sur les valeurs brutes de $\mathbf{y}$. Ceci se traduit notamment par une décroissance, du premier bloc au dernier, de la variance expliquée et des valeurs des coefficients PLS. Par conséquent, il est nécessaire de tester les différentes combinaisons possibles d'ordre des blocs.

3.5 La comparaison de modèles

Le pouvoir prédictif des étalonnages multivariés est généralement évalué par des critères statistiques tels que la RMSEP (Annexe A.5). Dans de nombreuses études, lors de l'optimisation d'un étalonnage multivarié, les comparaisons de modèles ne sont basées que sur les valeurs de RMSEP. Le modèle le plus performant est alors choisi en terme de RMSEP la plus faible. Cependant, lors de l'optimisation d'un étalonnage multivarié, les valeurs de RMSEP obtenues par différentes approches peuvent être relativement proches. Ainsi, il est souvent difficile de déterminer, sur la base de ces valeurs de RMSEP, si le

pouvoir prédictif des modèles est significativement différent.

Afin de procéder à une comparaison rigoureuse des performances des modèles de prédiction, un test statistique peut être réalisé. De nombreuses méthodes peuvent être mises en œuvre pour la comparaison statistique de modèles tels que le test de FISHER [49, 90], le "WILCOXON signed rank test" [91] et le "randomisation t-test" [103]. Toutes ces approches sont potentiellement utilisables mais également critiquables, rigoureusement parlé, car il n'existe pas de procédures établies pour la comparaison d'étalonnages multivariés. Dans cette partie, nous décrivons un test de comparaison de la distribution des erreurs de prédiction : le "Randomisation t-test". Nous discuterons également de l'intérêt des méthodes de *bootstrap* pour l'estimation de régions de confiance sur les prédictions des valeurs de propriétés. Nous parlerons ici de région de confiance, et non d'intervalle de confiance, car le bootstrap ne permet d'estimer que l'erreur engendrée par le modèle et non l'erreur totale sur la mesure. En effet, afin d'estimer l'erreur totale sur la mesure, il serait également nécessaire de tenir compte des erreurs effectuées sur les valeurs de référence et sur l'acquisition des spectres.

3.5.1 Le Randomisation t-test

Le "Randomisation t-test" a été introduit par VAN DER VOET [103]. Nous comparons deux modèles quantitatifs A et B pour la prédiction d'une propriété $\mathbf{y}$. Pour illustrer ce test, nous considérons ici que le lot de prédiction qui est composé de n échantillons dont les valeurs de référence de la propriété sont notées y_j (avec j=1, ..., n). Les valeurs prédites par les modèles A et B sont notés $\widehat{y}_{A,j}$ et $\widehat{y}_{B,j}$, respectivement. Les erreurs de prédiction des modèles A et B sont respectivement définies par $e_{A,j} = y_j - \widehat{y}_{A,j}$ et $e_B = y_j - \widehat{y}_{B,j}$. La comparaison des modèles A et B est effectuée sur les différences entre les erreurs de prédiction des modèles A et B. Pour cela, on défini $d_j = e^2_{A,j} - e^2_{B,j}$.

L'hypothèse nulle H_0 de ce test est : "Les erreurs quadratiques de prédiction des modèles A et B, $\mathbf{e}^2_A$ et $\mathbf{e}^2_B$, admettent la même distribution". Le test "randomisation t-test" est basé sur le fait que, si l'hypothèse H_0 est vérifiée, alors chaque valeur $|dj|$ est dérivée de dj ou $-dj$ avec une probabilité équivalente. Il en résulte qu'il est possible d'interchanger les valeurs $e^2_{A,j}$ et $e^2_{B,j}$ dans le calcul de dj sans en modifier la distribution.

Sur ce principe, le test procède donc comme suit. La procédure est également illustrée sur la Figure 3.12.

1. Calculer $d_j = e_{A,j}^2 - e_{B,j}^2$
2. Calculer le $T_{critique} = \overline{d} = \frac{\sum_{j=1}^{n_p} d_j}{n}$
3. Pour $m = 1, \ldots, M$, répéter les étapes 3a-3b
   (a) Générer une population $d_j(m)$ en attachant aléatoirement un signe à chaque d_j
   (b) Calculer $T_{observé}(m) = \overline{d}(m)$
4. Calculer la valeur de $p\text{-}value = \frac{k}{M+1}$

avec M : nombre d'itérations

k : nombre d'itérations pour lesquelles $\mathrm{T}_{observé} \geq \mathrm{T}_{critique}$

Pour que l'hypothèse nulle soit rejetée, il faut que la valeur de *p-value* soit inférieure ou égale au risque que l'on a choisi. Généralement, les tests statistiques sont calculés avec un risque de 5%. Ainsi, si la valeur de *p-value* est inférieure à 0,05, l'hypothèse nulle est rejetée. Il faut noter que la valeur de *p-value* ne peut jamais être inférieure à $\frac{1}{M+1}$. Ainsi, le nombre minimal d'itérations est M=19 pour que la valeur de *p-value* puisse atteindre la valeur minimale de *p-value* est 0,05. Cependant, Van der Voet conseille une valeur minimale de M=199 itérations. Ainsi, la valeur minimale de *p-value* est 0,005.

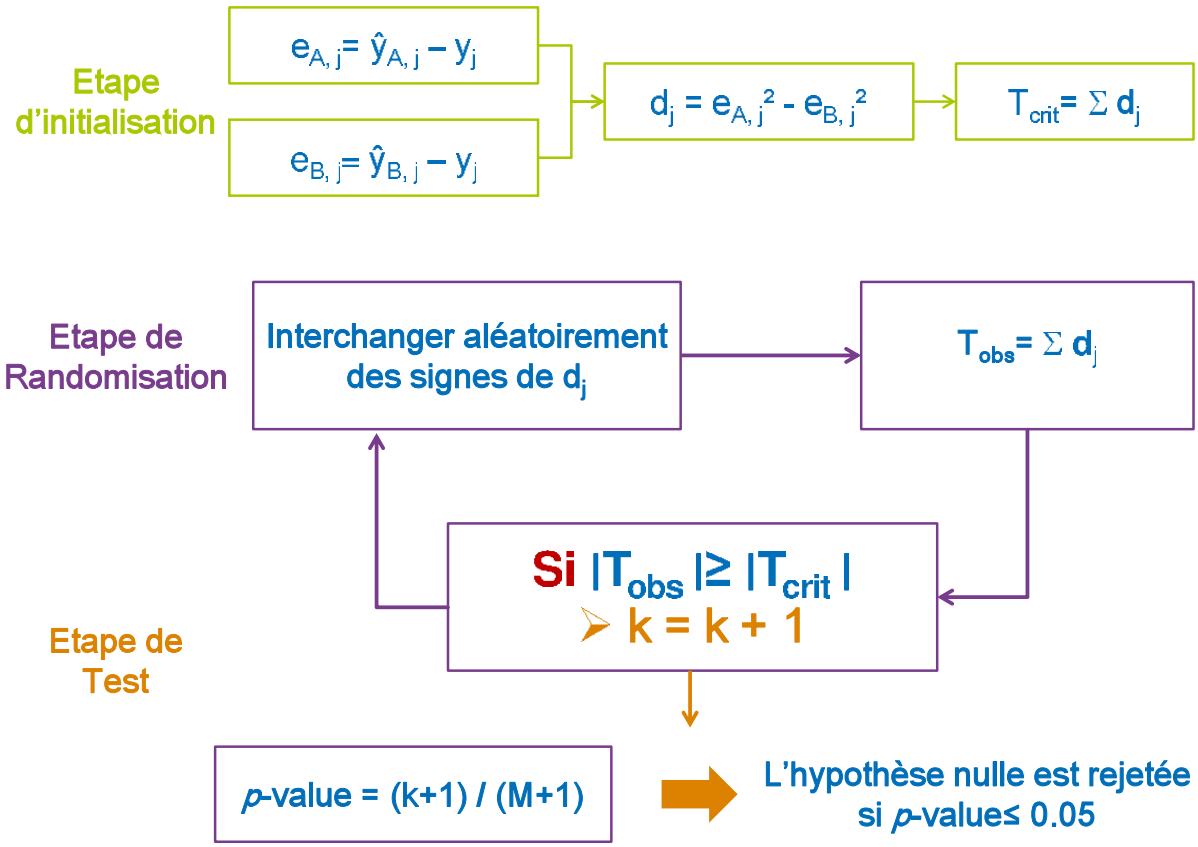


FIGURE 3.12 – Procédure du randomisation t-test

Afin d'illustrer ce test, prenons deux exemples extrêmes. Pour le premier exemple, $RMSEP_A > RMSEP_B$ et les erreurs quadratiques du modèle A, $e^2_{A,j}$, sont toutes très supérieures à celles du modèles B, $e^2_{B,j}$. Ainsi, les valeurs $d_j = e^2_{A,j} - e^2_{B,j}$ vont toujours être positives. La valeur du $T_{critique}$ sera alors grande. Lors de la procédure itérative, les valeurs de d_j, dont les erreurs quadratiques des modèles A et B auront été interchangées, vont être négatives. La valeur de $T_{observé}$ a donc de forte probabilité d'être inférieure à $|T_{critique}|$. La valeur de k sera donc faible et, par conséquent, la valeur de *p-value* sera inférieure à 0,05. L'hypothèse nulle sera rejetée. Il sera alors possible de conclure que les erreurs quadratiques de prédiction du modèle A et du modèle B ne suivent pas la même distribution et, par correspondance que la RMSEP du modèle A est significativement supérieure à celle du modèle B.

Pour le deuxième exemple, admettons que $RMSEP_A \approx RMSEP_B$, les erreurs quadra-

tiques du modèle A et celles du modèles B sont très proches. Les valeurs de d_j sont alors équitablement réparties autour de zéro (tantôt positives et tantôt négatives). Lors de la procédure itérative, lorsque les valeurs des erreurs quadratiques des modèles A et B vont être inter-changées, l'influence sur les valeurs de d_j va être très faible. Ainsi, les valeurs de $T_{observé}$ vont être distribuées autour de $T_{critique}$ et la valeur de k sera grande. La valeur de *p-value* sera donc supérieure à 0,05. Il sera alors possible de conclure que les erreurs quadratiques de prédiction du modèle A et du modèle B suivent la même distribution. Ainsi, nous pouvons conclure que les deux modèles prédictifs sont équivalents.

Le test "randomisation t-test" peut être appliqué pour la comparaison de modèle mais également pour fixer le nombre de facteurs dans le cadre de la régression PLS. Ce test a également été comparé à différents tests statistiques pour la comparaison de trois modèles avec $RMSEP_1 < RMSEP_2 < RMSEP_3$ [103]. Le test "randomisation t-test" a montré que les erreurs quadratiques du modèles 1 étaient significativement différentes de celles du modèle 2. Cependant, les résultats obtenus pour la comparaison des modèles 1 et 3 indiquaient qu'il n'y avait pas de différence significative entre les erreurs quadratiques de ces deux approches. Van Der Voet a interprété ce résultat par le fait que la haute valeur de la $RMSEP_3$ était due à deux échantillons très mal prédits. Les autres tests ont amené à la même conclusion sauf le test de Fisher. Cet exemple est très intéressant car il montre que le test "randomisation t-test" n'est pas sensible à quelques échantillons mal prédits mais bien à la distribution générale des erreurs de prédiction.

3.5.2 Les méthodes de *bootstrap*

Les méthodes de *bootstrap* ont été introduites par Efron [36] comme une méthode alternative pour l'estimation du biais et de la variance d'un estimateur, ainsi que pour la construction de régions de confiance. Cette méthode consiste à approcher la distribution d'une fonction statistique bâtie sur des observations. C'est une méthode de ré-échantillonnage basée sur des tirages aléatoires avec remise dans les données. Pour une introduction complète aux méthodes de *bootstrap*, nous pouvons citer le livre de Efron et Tibshirani [37]. Dans notre cas, le *bootstrap* est appliqué pour déterminer une région de confiance sur les valeurs prédites par les étalonnages multivariés. La technique *bootstrap*

appliquée dans ce travail est le *bootstrap* des résidus.

Dans notre cas, la procédure *bootstrap* consiste à générer une population, de taille B, de coefficient PLS β. Ainsi, il sera possible de prédire B fois les échantillons du lot de validation et ainsi obtenir une distribution des valeurs prédites par le modèle. Pour décrire la procédure, nous considérons le modèle PLS suivant :

$$\mathbf{y}_c = \mathbf{X}_c\beta + \mathbf{e}_c \tag{3.19}$$

avec $\mathbf{y}_c$: valeurs de propriétés des échantillons du lot d'étalonnage ($n_c \times 1$)

$\mathbf{X}_c$: absorbances des spectres des échantillons du lot d'étalonnage ($n_c \times m$)

β : coefficients PLS ($m \times 1$)

$\mathbf{e}_c$: résidus sur les valeurs prédites des échantillons du lot d'étalonnage ($n_c \times 1$)

Les coefficients PLS β et les résidus sont donnés par les Equations 3.20 et 3.21 respectivement.

$$\beta = \mathbf{W}(\mathbf{P}^T \cdot \mathbf{W})^{-1} \cdot \mathbf{q} \tag{3.20}$$

$$\mathbf{e}_c = \mathbf{y}_c - \mathbf{X}_c\beta \tag{3.21}$$

où $\mathbf{W}$ est la matrice des poids des variables, $\mathbf{P}$ et $\mathbf{q}$ sont les loadings des variables et de la réponse respectivement. Les valeurs de propriétés des échantillons de validation $\hat{y}_p$ sont prédites par le modèles PLS et la RMSEP est calculée suivant les Equations 3.22 et 3.23.

$$\hat{y}_p = X_p \cdot \beta \tag{3.22}$$

$$RMSEP = \sqrt{\frac{\sum_{j=1}^{n_p} \left(\hat{y}_{p,j} - y_{p,j}\right)^2}{n_p}} \tag{3.23}$$

Lorsque le modèle PLS est ainsi calculé, la technique *bootstrap* est appliquée. Le modèle théorique *bootstrap* est le suivant :

$$\mathbf{y}_c^* = \mathbf{X}_c\beta + \mathbf{e}_c^* \tag{3.24}$$

où β sont les coefficients PLS et $\mathbf{e}_c^*$ est un terme aléatoire issu des résidus $\mathbf{e}_c$ de la régression initiale. Nous décrivons, ci dessous, la procédure *bootstrap*, répétée B fois, pour obtenir ce modèle théorique. Les étapes de la procédure pour une itération b ($b = 1,\ldots, B$), également illustrées sur la Figure 3.13, sont les suivantes :

1. un tirage aléatoire avec remise dans la population initiale des résidus $\mathbf{e}_c = (\mathbf{e}_c)_{i=1\ldots n_c}$ est réalisé afin d'obtenir un échantillon *bootstrap* des résidus $\mathbf{e}_c^*(b) = (\mathbf{e}_c^*)_{i=1\ldots n_c}$;
2. un échantillon $\mathbf{y}_c^* = (\mathbf{y}_c^*)_{i=1\ldots n_c}$ est obtenu par : $\mathbf{y}_{c,i}^*(b) = \mathbf{y}_{c,i} + \mathbf{e}_{c,i}^*(b)$;
3. la régression PLS est appliquée pour déterminer $\beta^*(b)$ à partir de $\mathbf{X}_c$ et $\mathbf{y}_c^*(b)$;
4. les valeurs des échantillons de validation $\hat{y}_p^*(b)$ sont prédites à partir de $\mathbf{X}_v$ et $\beta^*(b)$;
5. la RMSEP(b) est calculée avec $\hat{y}_p^*(b)$ et $\mathbf{y}_p$.

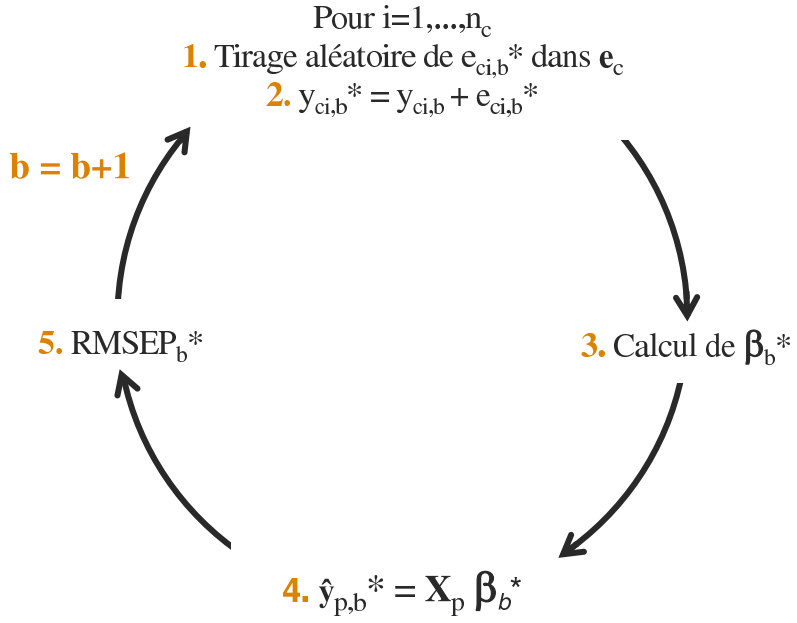


FIGURE 3.13 – Procédure *bootstrap* pour la génération d'un échantillon *bootstrap*

Une distribution G_i^*, de taille B, des valeurs prédites des échantillons de validation et des valeurs de RMSEP est donc obtenue. Les régions de confiance de chaque prédiction $\hat{y}_{p,i}$, au risque α et au degré de liberté ν, sont ensuite calculées. Si la distribution des valeurs $\hat{y}_p^*$ est :

- normale : la région de confiance est donnée par l'intervalle de prédiction standard : $\left[\hat{y}_{p,i} - s_{(\hat{y}_{p,i}^*)} \times t_{(1-\alpha)}^{\nu} \; ; \; \hat{y}_{p,i} + s_{(\hat{y}_{p,i}^*)} \times t_{(1-\alpha)}^{\nu}\right]$ où $s_{(\hat{y}_{p,i}^*)}$ est l'écart-type estimé de la distribution G_i^* des $\hat{y}_{p,i}^*$
- symétrique : la région de confiance est donnée par l'intervalle de confiance centile [1] : $[\hat{y}_{p,i} - G_i^*(\alpha/2) \; ; \; \hat{y}_{p,i} + G_i^*(1-\alpha/2)]$ où $G_i^*(\alpha/2)$ et $G_i^*(1-\alpha/2)$ représente les centiles, au risque α, de la distribution G_i^* ;
- "autre" : une des solutions est d'encadrer les valeurs prédites par les valeurs extrêmes de la distribution : $\left[\hat{y}_{p,i} - G_{i,min}^* \; ; \; \hat{y}_{p,i} + G_{(imax)}^*\right]$ où $G_{i,min}^*$ et $G_{i,max}^*$ représente respectivement les valeurs minimale et maximale de la distribution G_i^* ;

3.6 Bilan

La démarche analytique mise en œuvre pour le développement de l'analyse rapide des produits lourds a été présentée dans ce chapitre. Premièrement, nous avons décrit la base de données. Nous avons constaté que la base d'échantillons constituée est représentative des échantillons que l'on cherche à analyser en termes de gammes analytiques des propriétés, d'origines géographiques et de procédés. Nous avons également abordé que le nombre d'échantillons par propriété est globalement satisfaisant. Nous avons ensuite exposé l'instrumentation et les protocoles expérimentaux pour l'acquisition de spectres MIR en mode ATR et de spectres PIR en mode transmission.

Dans un deuxième temps, nous avons introduit les méthodes chimiométriques qui seront appliquées. En effet, nous avons exposé une méthode pour la co-optimisation par AG du choix des prétraitements et des variables à sélectionner. Les méthodes pour l'exploitation simultanée des spectres MIR et PIR ont ensuite été décrites. Enfin, nous avons

1. En statistiques, un centile est chacune des 99 valeurs qui divisent les données triées en 100 parts égales, de sorte que chaque partie représente 1/100 de l'échantillon de population. Le 1er centile sépare le 1% inférieur des données.

détaillé les procédures qui seront utilisées pour, d'une part, comparer les performances des différentes approches pour le calcul des modèles et, d'autre part, pour estimer un niveau de confiance sur les prédictions des valeurs prédites.

4 Développement de modèles d'analyse multivariée

Dans ce chapitre, nous présenterons les travaux réalisés pour le développement de modèles d'étalonnage multivarié. Nous nous concentrerons essentiellement ici sur l'optimisation du choix des méthodes de traitement des données et des techniques de régression. Il a été décidé de ne présenter que les résultats obtenus dans le cadre de la détermination des teneurs en Saturés, en Aromatiques, en Résines et en Asphaltènes (SARA). Ce choix se justifie d'abord par le fait que les premiers résultats obtenus pour la prédiction de ces propriétés étaient très insatisfaisants. Or, la méthode de référence du fractionnement SARA est très longue (8 heures) et justifie presque à elle-seule la mise en œuvre d'une analyse rapide des produits lourds. Ainsi, l'amélioration des modèles de prédiction de ces propriétés est un point important de ce travail de thèse. De plus, le fractionnement SARA consiste à séparer les molécules des produits lourds en fonction de leur polarité et donc, de leur degré d'insaturation. Ces propriétés permettent ainsi d'évaluer l'aptitude des modèles chimiométriques à caractériser des informations chimiques spécifiques et ont donc un fort potentiel d'interprétation pour faire le lien entre les modèles chimiométriques, l'information spectrale et l'information chimique.

Nous procéderons tout d'abord à l'exploration des données disponibles pour la caractérisation de la SARA, aussi bien pour ce qui est des valeurs de référence que pour les données spectrales. La démarche et les résultats de l'optimisation simultanée du choix du prétraitement des données spectrales et des variables à sélectionner par algorithmes génétiques seront ensuite présentés. Nous montrerons le potentiel de cette approche pour la définition d'une méthode de prétraitement adaptée ainsi que l'apport de la sélection de

variables pour l'interprétation globale des modèles. Enfin, l'étude de comparaison et de fusion des spectroscopies MIR et PIR sera décrite. Nous présenterons les améliorations obtenues pour certaines propriétés.

4.1 Exploration des données

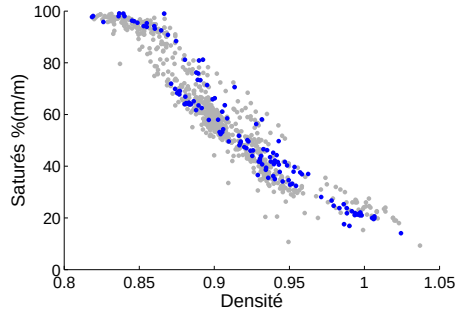
Nous présenterons tout d'abord les valeurs de référence disponibles pour la détermination de la SARA. L'interprétation des spectres sera ensuite réalisée. Les caractéristiques des bases spectrales MIR et PIR seront également détaillées. Enfin, une étude préliminaire de ces bases spectrales sera effectuée par une analyse en composantes principales (ACP).

4.1.1 Présentation des caractéristiques de la base d'échantillons SARA

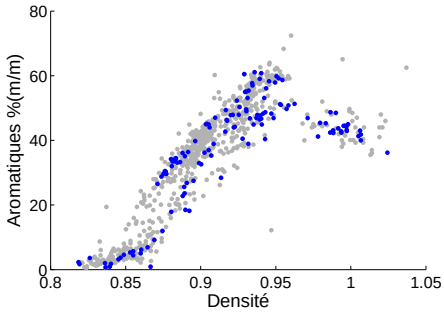
Il convient de rappeler que la base d'échantillons se compose de 230 distillats sous-vide (DSV) et résidus atmosphériques (RA). Sur ces 230 échantillons, la détermination des teneurs en SARA est disponible pour 133 échantillons. La Figure 4.1 représente les valeurs des échantillons de la base disponibles pour chaque fraction en fonction de la densité (en bleu). Les valeurs des échantillons d'archive ont également été indiquées (en gris) afin de déterminer si les échantillons de la base couvrent la gamme analytique de chaque propriété.

La Figure 4.1a illustre le fait que la teneur en composés saturés est inversement proportionnelle à la valeur de densité. Cette figure révèle que les échantillons de la base recouvrent bien la gamme analytique de la teneur en saturés définie par les échantillons d'archive. Sur la Figure 4.1b, nous pouvons également constater que la gamme analytique de la teneur en aromatiques est bien représentée par les échantillons de la base. Cette figure révèle aussi une diminution de la teneur en aromatiques lorsque la valeur de densité est supérieure à environ 0,98. La Figure 4.1c illustre la gamme analytique de la teneur en résines. Ici, on observe une forte augmentation de la teneur en résines à partir d'une valeur de densité de 0,98. Cette figure montre que l'échantillon entouré en rouge a une teneur en résines très supérieure aux autres échantillons de la base. Une décision sur la

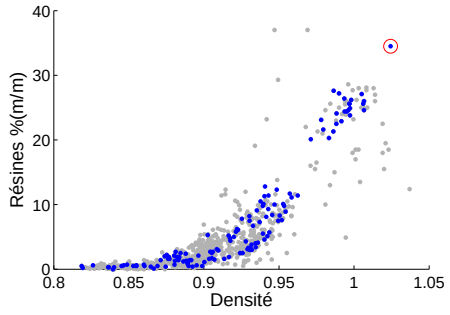
considération de cet échantillon sera prise lors du calcul de l'équation d'étalonnage. Enfin, la Figure 4.1d montre la gamme analytique de la teneur en asphaltènes. Nous pouvons constater que les teneurs en asphaltènes sont égales à zéro jusqu'à une valeur de densité qui se trouve ici aussi aux alentours de 0,98.



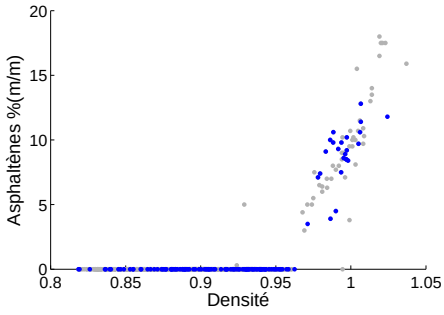
(a) Gamme analytique de la teneur en saturés



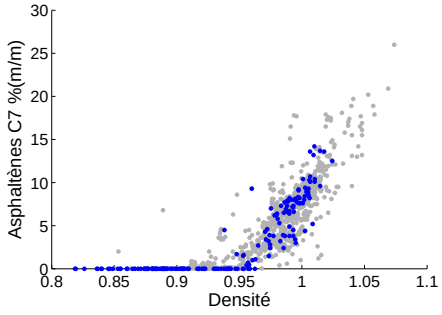
(b) Gamme analytique de la teneur en aromatiques



(c) Gamme analytique de la teneur en résines



(d) Gamme analytique de la teneur en asphaltènes



(e) Gamme analytique de la teneur en asphaltènes C7

FIGURE 4.1 – Gammes analytiques des teneurs en SARA - Echantillons de la base (en bleu) - Echantillons d'archive (en gris)

On peut donc en déduire que la teneur en résines subie une forte augmentation avec l'apparition d'asphaltènes dans les échantillons. De plus, lorsque la teneur en asphaltènes augmente, la teneur en aromatiques diminue et celle en résines augmente.

Sur la Figure 4.1d, on remarque la présence d'échantillons dont les teneurs en asphaltènes sont égales à zéro. Nous rappelons que ces échantillons sont des DSV. De manière générale, les méthodes quantitatives ne sont pas capables de prédire des teneurs égales à zéro. Le modèle de prédiction des asphaltènes ne concernera donc pas ces échantillons particuliers. Pour identifier ces échantillons qui ne contiennent pas d'asphaltènes, une méthode de classification sera donc utilisée. Elle n'est pas présentée dans ce manuscrit. Sur cette même figure, nous pouvons constater que seule une petite partie des échantillons a une teneur en asphaltènes supérieure à zéro (24 échantillons sur 133). Ces échantillons correspondent aux RA. Ceci s'explique par le fait que le fractionnement SARA est le plus souvent réalisé sur les coupes DSV et RSV. Le nombre d'échantillons disponibles pour la teneur en asphaltènes ne sera alors pas suffisant pour développer un étalonnage multivarié. Nous avons donc décidé de réaliser l'étude sur la teneur en asphaltènes C7 (Figure 4.1e) car le nombre de valeurs disponibles et non-nulles est plus important (91 échantillons). Les méthodes d'analyses sont différentes pour la détermination de la teneur en asphaltènes C7 et en asphaltènes obtenue par le fractionnement SARA. En effet, la teneur en asphaltènes contient les molécules insolubles dans le toluène à chaud contrairement à la teneur en asphaltènes C7 (Partie 1.3.2). Il existe donc un biais entre les valeurs obtenues par ces deux méthodes. Les caractéristiques de ces deux propriétés ne sont néanmoins pas foncièrement différentes.

4.1.2 Interprétation des spectres

Avant tout, il convient de rappeler que l'acquisition de spectres MIR des échantillons des produits lourds a été réalisée en mode réflexion totale atténuée (ATR). En effet, ce mode d'échantillonnage permet l'obtention de spectres MIR des produits lourds, qui sont très absorbants dans ce domaine, sans saturation du signal (Partie 2.1.1). La Figure 4.2 représente les spectres MIR de deux échantillons représentatifs de la base : un distillat sous-vide (DSV) et un résidu atmosphérique (RA).

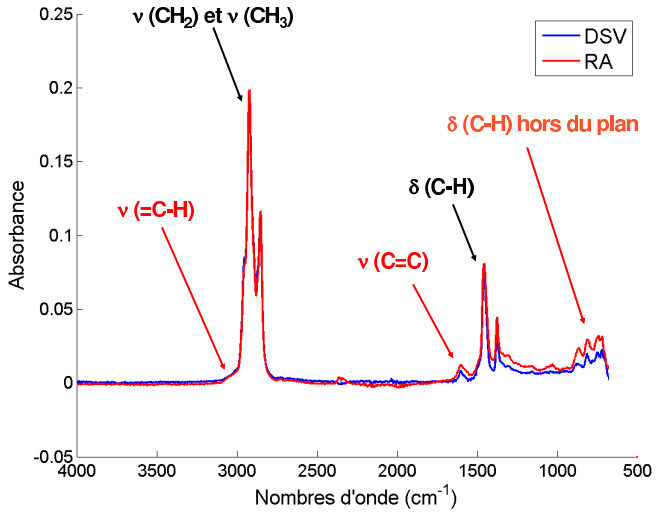


FIGURE 4.2 – Spectres MIR d'un DSV et d'un RA. L'interprétation des bandes d'absorption est indiquée sur les spectres (δ : vibration de déformation et ν : vibration d'élongation)

Cette figure illustre que les spectres MIR des échantillons DSV et RA sont très similaires et qu'ils présentent les mêmes bandes, dont l'interprétation est détaillée ci-dessous :

- **950-700 cm^{-1}** : déformations des liaisons C-H hors du plan (δ_{CH} hors du plan). Ces vibrations peuvent être attribuées à des cycles aromatiques ;
- **1500-1300 cm^{-1}** : déformations des liaisons C-H (δ_{CH}) ;
- **1650-1550 cm^{-1}** : élongations des doubles liaisons carbone-carbone ($\nu_{C=C}$) ;
- **3000-2750 cm^{-1}** : élongations symétriques et asymétriques des liaisons C-H dans les groupements CH_2 (2922 et 2852 cm^{-1}) et CH_3 (2953 cm^{-1}) (ν_{CH_2} et ν_{CH_3}) ;
- **3100-3000 cm^{-1}** : faible bande d'élongation des liaisons =C-H ($\nu_{=C-H}$).

Nous rappelons également que les spectres PIR des produits lourds ont été enregistrés en mode transmission. Nous avons opté pour ce mode d'échantillonnage car il permet d'obtenir une meilleure fidélité sur la mesure spectrale qu'en mode réflexion diffuse (Partie 2.1.2). De plus, les produits lourds du pétrole peuvent présenter une forte hétérogénéité. Le mode transmission permet donc de limiter les problèmes de représentativité de l'analyse. Enfin, le spectromètre PIR utilisé (Partie 3.6) présente de nombreux avantages :

- un trajet optique très faible (500 μm) qui permet d'éviter la saturation du signal malgré la forte absorption de ces produits ;
- une cellule unique, ce qui évite les variations de la longueur du trajet optique ;

- un protocole expérimental spécialement conçu pour l'échantillonnage de ces produits très visqueux dans la cellule ;
- une cellule thermostatée pour éviter, d'une part, les variations de température pendant l'acquisition du spectre et, d'autre part, le bouchage du circuit par solidification de l'échantillon.

La Figure 4.3 représente les spectres PIR de ces mêmes échantillons. Cette figure illustre que les échantillons DSV et RA présentent des bandes communes :

- **4460-4000 cm^{-1}** : combinaison entre les élongations symétriques et asymétriques des liaisons C-H dans les groupements CH_2 et CH_3 et les déformations des liaisons C-H ;
- **4700-4540 cm^{-1}** : combinaison entre les élongations des doubles liaisons C=C et les élongations des liaisons =C-H aromatiques ;
- **6200-5400 cm^{-1}** : première harmonique des élongations symétriques et asymétriques des groupements CH_2 et CH_3 ;
- **8600-8000 cm^{-1}** : seconde harmonique des élongations symétriques et asymétriques des groupements CH_2 et CH_3.

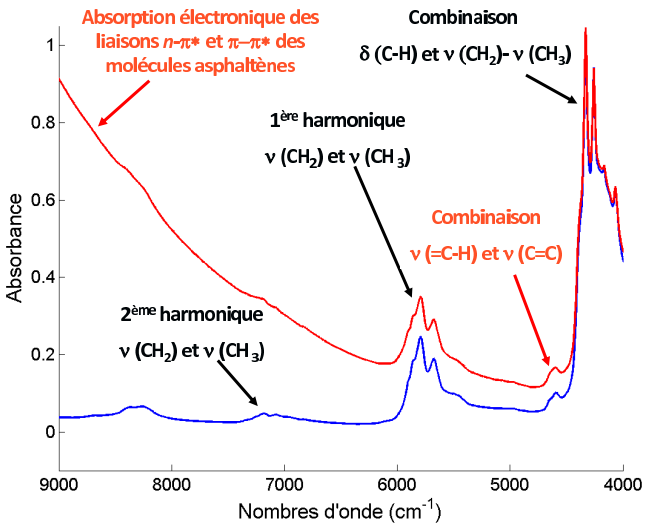


FIGURE 4.3 – Spectres PIR d'un DSV et d'un RA. L'interprétation des bandes d'absorption est indiquée sur les spectres (δ : vibration de déformation et ν : vibration d'élongation)

Sur le spectre du RA (en rouge), nous pouvons observer entre 9000 et 6000 cm^{-1} une forte déformation de la ligne de base lorsque l'on se déplace vers les nombres d'onde élevés. Nous désignerons ce phénomène par "une courbure de la ligne de base". MULLINS [80] a attribué cet effet à une absorption causée par les transitions électroniques des liaisons $n - \pi^*$ et $\pi - \pi^*$ des agrégats d'asphaltènes. Pour ce faire, MULLINS a procédé à l'acquisition des spectres PIR d'un bitume et de ces fractions maltènes [1] et asphaltènes. Il a remarqué que la courbure de la ligne de base est présente sur le spectre du bitume et des asphaltènes. Cependant, le spectre des maltènes ne présente pas cet effet.

Afin de définir si ce phénomène correspond à une diffusion ou une absorption, MULLINS a dilué un pétrole brut contenant des asphaltènes dans du tétrachlorure de carbone (CCl_4). En effet, le CCl_4 est connu pour dissocier les agrégats d'asphaltènes. Il a ensuite enregistré les spectres PIR du produit brut dilué et non-dilué en corrigeant la dilution par une augmentation du trajet optique. Les résultats obtenus ont montré que cet effet est présent sur les spectres du pétrole brut dilué et non dilué. Or, la diffusion est très corrélée à la taille des particules. Ainsi, si cet effet était due à de la diffusion, la dissociation des agrégats d'asphaltènes aurait engendré une forte diminution de son intensité. Par conséquent, MULLINS a conclu que la courbure de la ligne de base est due à la queue de la bande d'absorption observée en spectroscopie dans le domaine du visible, causée par les transitions électroniques des liaisons $n - \pi^*$ et $\pi - \pi^*$ des agrégats d'asphaltènes.

L'interprétation des spectres MIR et PIR nous a permis de répertorier les principales bandes d'absorption. Nous disposons aussi d'informations sur la composition chimique des produits lourds et sur leurs propriétés de caractérisation (Chapitre 1). Nous avons donc mené une réflexion pour essayer de définir quelles sont les interactions entre ces différentes informations.

Les bandes d'absorption présentent dans les spectres sont majoritairement dues aux liaisons contenant des atomes de carbone et d'hydrogène. L'information nécessaire à la détermination des teneurs en carbone, en carbones insaturés et en hydrogène est donc présente. De plus, une absorption électronique des liaisons $n - \pi^*$ et $\pi - \pi^*$ causée par les agrégats d'asphaltènes est présente dans les spectres PIR. Il est donc envisageable de

1. Les maltènes correspondent au produit désasphalté. Ils contiennent donc les saturés, les aromatiques et les résines.

pouvoir relier cette information à la teneur en asphaltènes. L'analyse de la SARA est une analyse par famille en fonction de la polarité et donc, en partie, du degré de saturation. Nous avons observé que les bandes sont dues à des groupements de degrés de saturation différents (groupement CH_2 et CH_3 , liaison C=C, cycles aromatiques...). La prédiction des teneurs en saturés, aromatiques et résines à partir des spectres de vibration semble donc envisageable. Enfin, les propriétés physico-chimiques globales ne sont pas reliées à une famille chimique spécifique mais font état de la matrice globale du produit. Or, les spectres obtenus en spectroscopie vibrationnelle sont décrits comme des "empreintes spectrales" du produit. Les variations des valeurs de propriétés, comme la densité ou la viscosité, qui sont reliées à des variations globales de la composition du produit devraient être potentiellement détectées en spectroscopie vibrationnelle. Lorsque la propriété varie linéairement avec la composition, il est possible de relier, sans trop de difficultés, ces variations spectrales aux valeurs des propriétés. Par contre, cela peut-être plus difficile pour des propriétés comme la viscosité dont les variations ne sont pas linéairement reliées à la composition. Les hétéroéléments (soufre, azote) et les métaux (nickel et vanadium) sont présents à de très faibles teneurs (de quelques pour-cent pour le soufre à quelques dizaines de ppm pour le nickel). De plus, aucune information spécifique à ces éléments n'a été observée parmi les principales bandes des spectres MIR et PIR. L'analyse par spectroscopie vibrationnelle des hétéroéléments et des métaux paraît alors difficile car, comme nous l'avons évoqué dans la Partie 2.3.3, la spectroscopie vibrationnelle n'est généralement pas adaptée aux analyses de traces.

4.1.3 Présentation des bases spectrales

Concernant les spectres MIR bruts des échantillons de la base (Figure 4.4), nous pouvons observer les vibrations d'élongation (4000-3500 cm^{-1}) et de déformation (1700-1400 cm^{-1}) des liaisons O-H des molécules d'eau. De plus, à 2300 cm^{-1}, nous pouvons constater la présence d'une bande correspondant à l'élongation des liaisons C-O du dioxyde de carbone. Ces bandes sont dues aux variations de la teneur de ces molécules dans l'air ambiant.

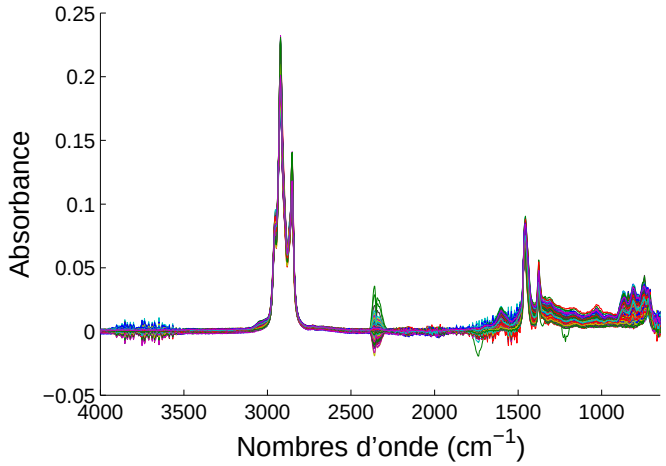


FIGURE 4.4 – Spectres MIR bruts des échantillons de la base

Les variations sur le domaine 4000-3500 cm^{-1} et à 2300 cm^{-1} ne sont pas problématiques car il n'y a pas de bande d'absorption significative sur ces gammes spectrales. Ces zones pourront donc être retirées du spectre lors du développement des modèles d'étalonnage multivarié. En revanche, les bandes sur la gamme 1700-1400 cm^{-1} interfèrent avec la bande due aux vibrations d'élongation des doubles liaisons carbone-carbone (1650-1550 cm^{-1}). Ce phénomène peut dégrader la qualité des modèles. Nous avons utilisé des méthodes de lissage pour réduire ces variations. Les résultats obtenus ont permis de réduire ces interférences. L'impact sur le pouvoir prédictif et la robustesse des modèles est néanmoins faible. Nous avons donc opté pour la présentation des résultats sans l'application de ces techniques de lissage.

Si l'on s'intéresse maintenant aux spectres PIR de la base, nous pouvons tout d'abord noter la présence d'un spectre dont le signal est saturé entre 9000 et 8000 cm^{-1} (Figure 4.5). Il ne sera donc pas considéré lors du développement des modèles. Ce spectre correspond à un échantillon dont la densité et la teneur en asphaltènes sont parmi les plus élevées. Il illustre également que, lorsque les valeurs d'absorbance sont élevées (à partir de 2 unités d'absorbance), le signal est bruité. Ce phénomène peut se retrouver sur une dizaine de spectres caractérisés par une forte absorption électronique. Pour les mêmes raisons que mentionné précédemment, nous présenterons les résultats sans l'application préalable d'une technique de lissage. Sur cette figure, nous pouvons également remarquer que l'on peut différencier deux lots de spectres. Les spectres qui admettent une absorption

électronique (coupe RA) et les spectres dont la ligne de base ne varie pas (coupe DSV). A partir de ce constat, il est légitime d'envisager de séparer la base en deux lots : un lot pour les échantillons DSV et un lot pour les échantillons RA.

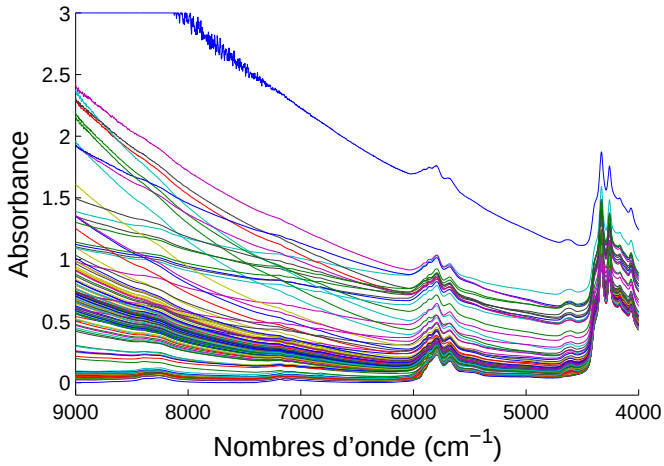


FIGURE 4.5 – Spectres PIR bruts des échantillons de la base

Le Tableau 4.1 montre que les gammes analytiques de la plupart des propriétés des DSV et des RA se chevauchent. L'hétérogénéité entre ces deux coupes pétrolières n'est donc pas due à des compositions chimiques distinctes mais à une signature spectrale différente engendrée par la présence d'asphaltènes dans les RA. Par conséquent, la séparation de la base de données provoquerait une diminution du nombre d'échantillons disponibles pour développer les modèles prédictifs. Nous avons donc choisi de développer des modèles d'analyse multivariée communs à ces deux coupes pétrolières, ce qui nécessitera d'ailleurs un traitement des données spectrales adapté et optimisé.

Tableau 4.1 – Gamme analytique en fonction de la coupe pétrolière

Propriété	**Gamme analytique**	
	DSV	**RA**
Densité	0,8186 - 0,9625	0,853 - 1,0243
Viscosité cinématique à 100°C (cSt)	3,45 - 14,58	5,45 - 793,19
Saturés %(m/m)	32,4 - 99,1	14,1 - 99,0
Aromatiques %(m/m)	0,7 - 61,1	0,9 - 48,7
Résines %(m/m)	0 - 12,8	0,1 - 34,5
Asphaltènes %(m/m)	0	0 - 12,8
Carbone Conradson %(m/m)	0,05 - 0,6	2,33 - 20,03
Asphaltènes C7 %(m/m)	0	0 - 14,2
Carbone %(m/m)	86,12 - 88,40	86,01 - 87,90
Hydrogène %(m/m)	10,83 - 14,59	9,95 - 14,04
Soufre %(m/m)	1.10^{-4} - 2,52	9.10^{-4} - 4,77
Azote (ppm)	0,1 - 4450,0	0,2 - 11 800,0
Carbone aromatique %(m/m)	2 - 34,3	22 - 36,8

4.1.4 Visualisation par ACP

Afin de visualiser les bases spectrales, une ACP sur les spectres a été réalisée. Pour la base spectrale MIR, les zones spectrales contenant les bandes dues aux variations des teneurs en eau et en dioxyde de carbone dans l'air ambiant (4000-3200 cm^{-1} et 2700-1700 cm^{-1}) ont été supprimées (Partie 4.1.2). En effet, l'introduction de ces variations dégraderait la pertinence de l'interprétation de l'ACP car elles ne sont pas reliées à la composition chimique. L'ACP a donc été calculée sur les spectres MIR prétraités en dérivée 1ère sur les domaines 3200-2700 cm^{-1} et 1700-650 cm^{-1} (Figure 4.6). L'ACP sur la base spectrale PIR a été calculée en utilisant la dérivée 1ère sur le domaine 9000-4000 cm^{-1} (Figure 4.7).

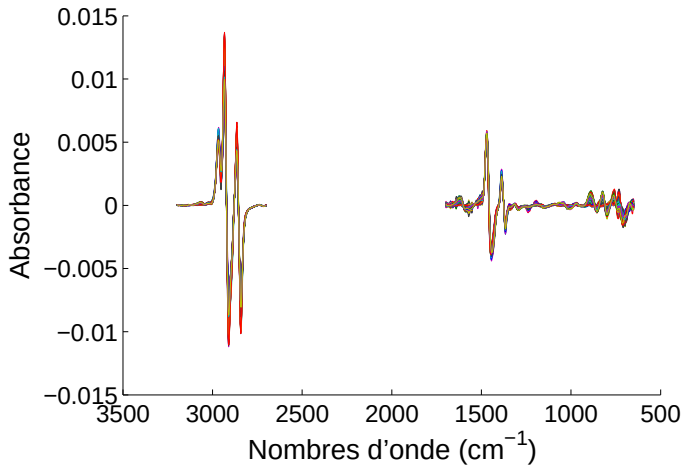


FIGURE 4.6 – Spectres MIR en dérivée 1ère sur le domaine 4000-3200 et 2700-1700 cm^{-1}

La Figure 4.8 synthétise les résultats de l'ACP sur les spectres MIR. La Figure 4.8b illustre les scores de l'ACP des spectres des échantillons de la base. Nous pouvons constater que la répartition des échantillons est homogène sur les trois premières composantes. De plus, la valeur de densité de chacun des échantillons est indiquée par un code couleur. Les Figures 4.8a et 4.8b illustrent que la distribution des échantillons est fonction de la densité. Ces trois premières composantes représentent environ 90% de la variance de la base spectrale MIR. Ce constat nous permet de confirmer que les variations spectrales sont reliées à la densité et, par conséquent, à la composition chimique des produits lourds.

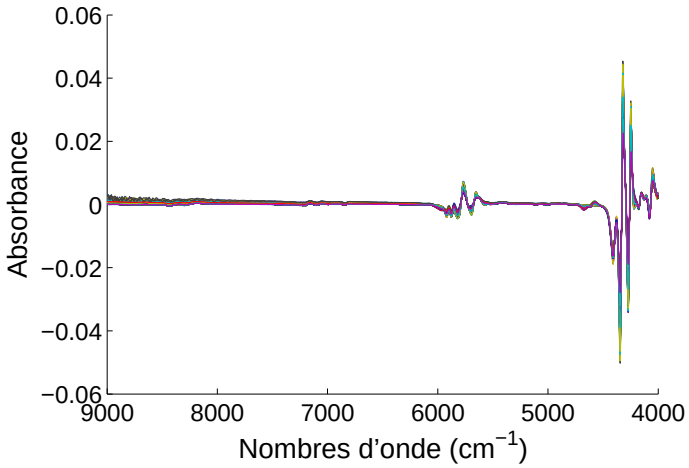
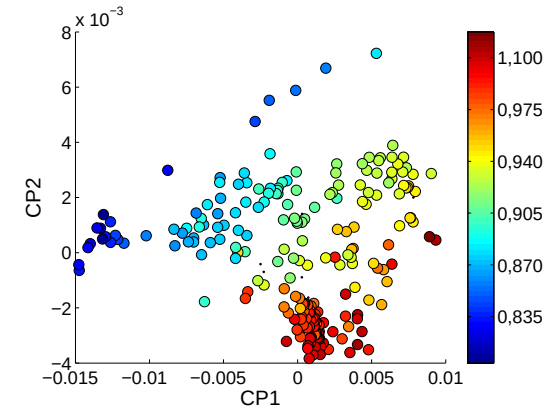
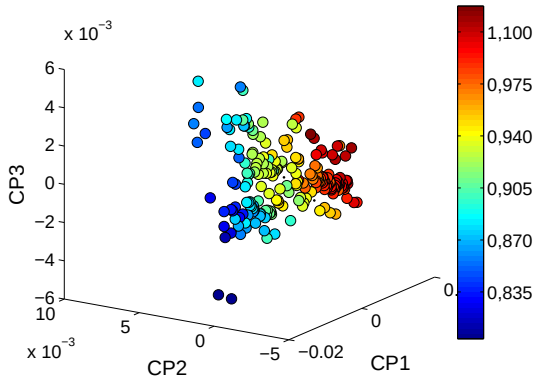


FIGURE 4.7 – Spectres PIR en dérivée 1ère sur le domaine 9000-4000 cm^{-1}

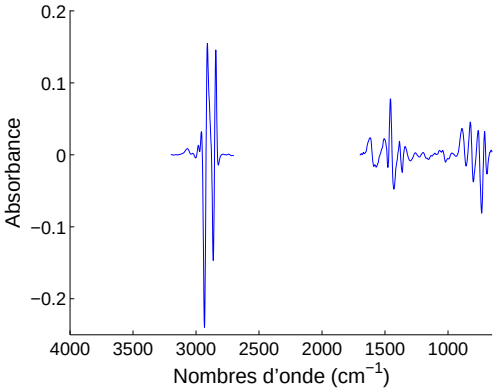
Les trois premières composantes expliquent respectivement 69,7 %, 13,9 % et 6,3 % de la variance expliquée. Les Figures 4.8c et 4.8e illustrent l'importance des bandes dues aux groupements saturés sur les composantes 1 et 3 : déformations des liaisons C-H (1500-1300 cm^{-1}) et élongation symétrique et asymétrique des liaisons C-H dans les groupements CH_2 et CH_3 (3000-2750 cm^{-1}). En revanche, nous observons une forte influence des bandes dues aux vibrations des liaisons dans les groupements insaturés sur la deuxième composante (Figure 4.8d) : bandes de déformation des liaisons C-H hors du plan attribuées aux cycles benzéniques (950-700 cm^{-1}), l'élongation des doubles liaisons C=C (1650-1550 cm^{-1}) et la bande due aux élongations des liaisons =C-H (3100-3000cm^{-1}). En effet, ces bandes ont une intensité relativement faible dans les spectres MIR des produits lourds (Figure 4.6). Or, leur intensité relative par rapport aux bandes dues aux groupements saturés est importante sur cette composante.



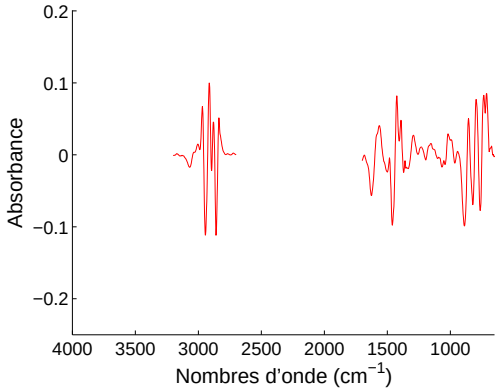
(a) Scores sur les deux premières composantes en fonction des valeurs de densité



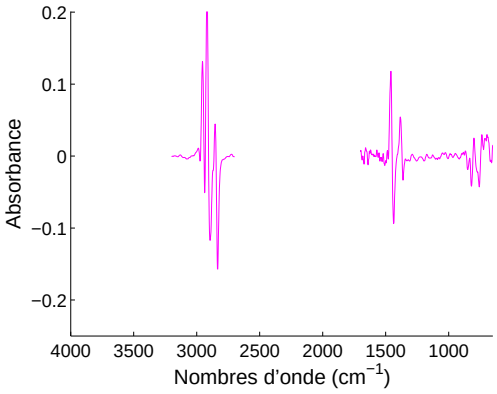
(b) Scores sur les trois premières composantes en fonction des valeurs de densité



(c) Loadings de la première composante (69,7% de variance expliquée)



(d) Loadings de la deuxième composante (13,9% de variance expliquée)



(e) Loadings de la troisième composante (6,3% de variance expliquée)

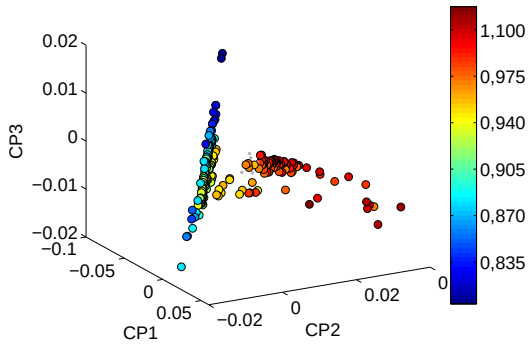
FIGURE 4.8 – Synthèse des résultats de l'ACP sur les spectres MIR

La Figure 4.9 synthétise les résultats de l'ACP sur les spectres PIR. Deux constats peuvent être faits sur les représentations des scores des échantillons (Figures 4.9a et 4.9b).

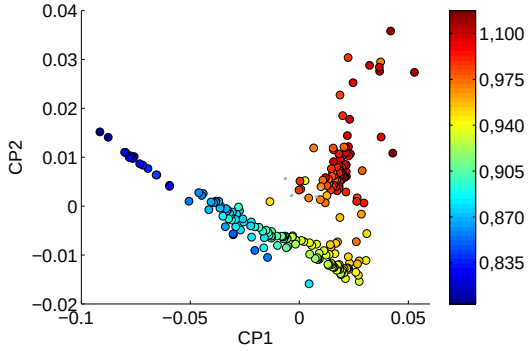
Premièrement, les trois premières composantes expliquent 97% de la variabilité de la base spectrale PIR et la distribution des échantillons est relativement fonction des valeurs de densité. La variablilité des spectres est donc reliée aux valeurs de densité et, par conséquent, à la composition chimique globale des produits lourds du pétrole.

Deuxièmement, la distribution des échantillons n'est pas uniforme sur les trois premières composantes. Les Figures 4.9a et 4.9b montrent la présence de deux groupes d'échantillons. En effet, on observe que ces deux groupes appartiennent chacun à un plan de l'espace et que ces deux plans sont quasi orthogonaux. Les teneurs en asphaltènes C7 sont indiquées sur la Figure 4.9c. On constate qu'un premier groupe d'échantillons (en bleu foncé) correspond aux produits lourds qui ne contiennent pas d'asphaltènes (coupe DSV) tandis que le deuxième (bleu clair à rouge foncé) correspond à ceux qui ont une teneur en asphaltènes supérieure à zéro (coupe RA). Sur la deuxième composante, nous pouvons constater que les coefficients augmentent lorsque l'on se déplace vers les nombres d'onde élevés. Nous en avons déduit que la deuxième composante explique une partie de la variabilité liée à l'absorption électronique. La dérivée 1$^{\text{ère}}$ ne corrige donc pas totalement la courbure de la ligne de base. En effet, la Figure 4.9e illustre qu'une variabilité persiste sur le domaine 9000-6000 cm^{-1}. Ceci explique alors la distribution des échantillons en deux lots en fonction de la teneur en asphaltènes.

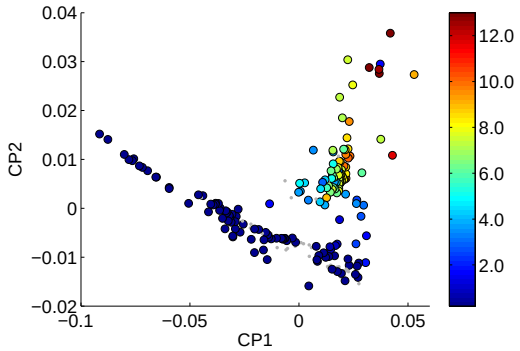
La dérivée seconde permet de supprimer totalement la courbure de la ligne de base car elle est capable de corriger à la fois les effets additifs et multiplicatifs présents dans les spectres des produits lourds (Partie A.1.1). Nous avons cependant évoqué le fait qu'une partie des spectres PIR ont une absorbance très élevée sur la gamme 9000-6000 cm^{-1}. Le signal transmis et, par conséquent, le rapport signal sur bruit sont donc faibles. Lorsque l'on augmente le degré de la dérivée, le rapport signal sur bruit est dégradé. Les étalonnages multivariés développés à partir des spectres prétraités par la dérivée 1$^{\text{ère}}$ ont donc conduit à de meilleures prédictions que ceux développés à partir des spectres prétraités par la dérivée 2$^{\text{nde}}$ (Partie 4.2.2.2). C'est pourquoi, nous avons choisi de présenter l'ACP sur les spectres prétraités par la dérivée 1$^{\text{ère}}$.



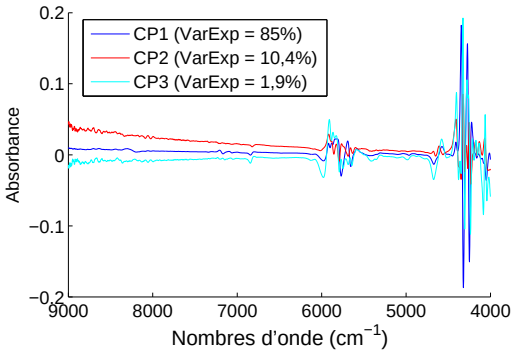
(a) Scores sur les trois premières composantes en fonction des valeurs de densité



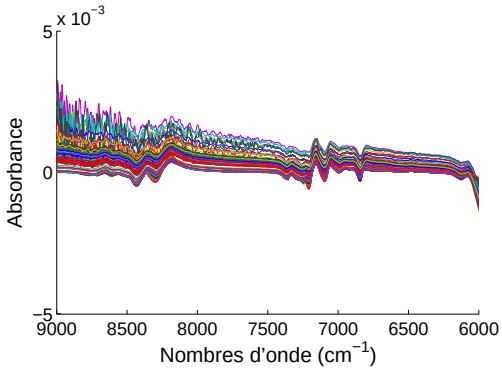
(b) Scores sur les deux premières composantes en fonction des valeurs de densité



(c) Scores sur les deux premières composantes en fonction de la teneur en asphaltènes C7



(d) Loadings des 3 premières composantes



(e) Spectres PIR dérivés entre 9000 et 6000 cm^{-1}

FIGURE 4.9 – Synthèse des résultats de l'ACP sur les spectres PIR

4.1.5 Bilan

Nous avons présenté dans cette partie les données qui vont servir au développement de modèles d'analyse multivariée. Nous avons tout d'abord décrit les caractéristiques des échantillons disponibles pour la prédiction de la teneur en SARA. Nous avons ensuite procédé à l'interprétation des bandes spectrales. On peut notamment noter qu'une absorption électronique des liaisons $n-\pi^*$ et $\pi-\pi^*$ causée par la présence d'agrégats d'asphaltènes a été observée en spectroscopie PIR. Cette absorption se traduit par une forte augmentation des valeurs d'absorbance quand on se déplace vers les hauts nombres d'onde.

La description par ACP des spectres MIR de la base d'échantillons a montré que la répartition des échantillons est homogène. En revanche, il a été constaté que les spectres PIR se divisent en deux lots d'échantillons. La présence de ces deux groupes d'échantillons nécessitera un traitement adapté des spectres PIR pour les prédictions des teneurs en saturés, en aromatiques et en résines. Dans le cas de la détermination de la teneur en asphaltènes C7, l'absorption électronique peut s'avérer très utile. Les variations de cette absorption sont néanmoins trop importantes. Il faudra alors les réduire sans les éliminer. Ainsi, nous avons décidé d'évaluer l'apport d'un algorithme d'optimisation, les algorithmes génétiques (AG), pour le choix du prétraitement et de la sélection de variables dans le cadre des développements des modèles de prédiction de ces quatre propriétés.

Enfin, les bases spectrales MIR et PIR ont des caractéristiques différentes et contiennent de l'information spécifique. Nous nous attendons donc à obtenir des performances inégales pour les modèles développés séparément à partir des spectroscopies MIR et PIR. Nous réaliserons donc une étude comparative des pouvoirs prédictifs des étalonnages multivariés calculés à partir de ces deux techniques spectroscopiques. Dans un deuxième temps, nous évaluerons le potentiel de l'exploitation simultanée des spectroscopies MIR et PIR (Partie 4.3).

4.2 Optimisation simultanée du prétraitement et de la sélection de variables des modèles PIR par algorithmes génétiques

Dans la Partie 3.3, nous avons évoqué que les données spectrales ne sont pas forcément sous la forme la plus adaptée au développement de l'équation d'étalonnage. Il est donc nécessaire, d'une part, d'appliquer des prétraitements pour éliminer les variations non-désirées. D'autre part, il faut identifier les variables pertinentes pour la description de la propriété considérée. Nous avons également mentionné la nécessité d'une optimisation du choix du prétraitement et des variables à sélectionner. A ce titre, deux approches sont possibles.

Premièrement, une démarche essai-erreur consistant à tester un nombre fini et souvent restreint de méthodes de prétraitement et de domaines spectraux peut-être utilisée. Les limitations de cette approche ont été exposées (Partie 3.3).

Deuxièmement, nous avons mentionné le potentiel de l'apport des méthodes de la sélection de variables. Un algorithme d'optimisation est généralement mis en œuvre pour cette approche. Nous avons évoqué dans la Partie 3.3 que les AG offrent un bon compromis entre l'exploration et l'exploitation par rapport aux autres méthodes d'optimisation. De plus, ils sont très adaptables à de nombreux problèmes d'optimisation et ils ont fait l'objet de nombreux travaux en chimiométrie pour l'optimisation de la sélection de variables dans le cadre de développement d'étalonnages multivariés [62, 64]. Généralement, le choix du prétraitement est effectué préalablement à l'application d'une méthode de sélection de variables. Cette démarche est justifiée par le fait que les prétraitements sont généralement appliqués pour corriger des variations dues à l'acquisition des spectres tels que les interférences aléatoires (bruit), physico-chimiques (diffusion) ou liées à l'instrumentation (non-linéarité du détecteur dans certaines gammes ou variation du trajet optique). La sélection de variables est ensuite appliquée pour identifier les variables pertinentes, c'est à dire l'information chimique nécessaire pour la description de la propriété. Nous pensons que l'application du prétraitement modifie également les variations des bandes spectrales,

c'est à dire l'information chimique contenue dans le spectre, et peut-donc influencer la sélection de variables. C'est pourquoi, nous avons opté pour l'évaluation d'une procédure d'optimisation par AG pour le choix simultané des méthodes de prétraitements et des variables à sélectionner [31].

Nous avons observé dans la partie précédente que la base spectrale PIR se divisait en deux groupes d'échantillons selon la présence ou l'absence de l'absorption électronique et, donc d'asphaltènes. Nous avons mentionné dans la Partie 4.1.1 que seul les échantillons RA, qui contiennent des asphaltènes sont considérés. Seul le groupe d'échantillons RA est donc pris en compte pour la détermination de la teneur en asphaltènes C7. Dans le cadre de l'établissement de ce modèle de prédiction de la teneur en asphaltènes C7, l'absorption électronique des liaisons $n - \pi^*$ et $\pi - \pi^*$ devrait donc être une information pertinente. La variabilité de cette absorption électronique est néanmoins très importante et devra probablement être réduite sans être éliminée. L'interaction entre le(s) prétraitement(s) appliqué(s) et les variables sélectionnées est donc très importante dans le cadre de ce modèle. C'est pourquoi nous avons décidé d'appliquer la procédure d'optimisation simultanée pour le développement de cet étalonnage multivarié.

De plus, les deux groupes d'échantillons sont considérés pour le développement des modèles de prédiction des teneurs en saturés, en aromatiques et en résines. La présence de deux groupes d'échantillons peut alors être problématique pour le calcul de ces étalonnages multivariés et conduire à des performances non-optimales. Ainsi, nous avons décidé d'évaluer le potentiel de la procédure d'optimisation par AG à proposer une correction de cette distribution des échantillons.

Dans un premier temps, nous décrirons la démarche utilisée pour l'optimisation des modèles de prédiction à partir des spectres PIR. Nous exposerons et discuterons par la suite les résultats obtenus.

4.2.1 Démarche de l'optimisation par AG

Nous commencerons ici par exposer les paramètres de configuration utilisés pour l'optimisation par AG. Nous décrirons ensuite la démarche appliquée pour l'optimisation du prétraitement et de la sélection de variables de chaque propriété ainsi que la procédure

pour la détermination du modèle final. Enfin, nous décrirons l'approche mise en place pour l'évaluation des performances des modèles et l'interprétation des variables sélectionnées.

4.2.1.1 Paramètres de configuration des AG

Les paramètres utilisés pour l'optimisation simultanée du prétraitement et de la sélection de variables sont résumés dans le Tableau 4.2.

Les calculs d'optimisation par AG sont effectués sur un cluster informatique. Il est composé d'unités de calcul autonomes (un ordinateur principal et 16 processeurs) qui sont reliées entre elles à l'aide d'un réseau de communication. Dans le cadre de l'optimisation par AG, les différentes unités servent aux calculs de la fonction d'adaptation qui sont généralement les plus chronophages. Par conséquent, le temps de calcul n'est pas très important. Ainsi, une population de 256 individus a pu être utilisée. Cette population de taille importante permet une meilleure exploration de l'espace des solutions potentielles.

Lors de la création de la population initiale, la sélection des prétraitements est effectuée de manière aléatoire. D'autre part, les variables du spectre sont regroupées par groupes de 20, ce qui représente des intervalles spectraux de 38 cm^{-1}. Ceci permet de réduire le nombre de solutions potentielles et donc de permettre une convergence plus rapide de l'algorithme. Enfin, seules 10 % des variables de chaque chromosome peuvent être sélectionnées à l'initialisation. En effet, un nombre trop important de variables sélectionnées ralenti fortement la convergence de l'algorithme.

Dans le cadre de l'optimisation de modèles de prédiction, la fonction d'adaptation correspond à la minimisation de l'erreur de prédiction, qui est évaluée par validation croisée. Nous avons ici opté pour une validation croisée partielle (Annexe A.4). Ainsi, 10 itérations, qui correspondent chacune au tirage avec remise d'un groupe d'échantillons, sont réalisées pour chaque chromosome. Cette méthode de validation croisée permet une évaluation plus juste, c'est à dire moins optimiste, que la validation croisée totale [6, 26]. Cette méthode réduit alors le risque de sur-ajustement. Le nombre de facteurs PLS est déterminé automatiquement lors de la procédure d'optimisation. Le nombre de facteurs est fixé à A lorsque l'amélioration de la RMSECV entre A et A+1 est inférieure à 5%.

Pour la création de la nouvelle population, un pourcentage de recombinaison de 50% est

Tableau 4.2 – Paramètres des AG pour la co-optimisation

Étapes de la procédure	**Paramètres**	**Valeur**
Création de la population initiale	Taille de la population	256 individus
	Nombre p de prétraitements	1 à 4
	Sélection des prétraitements	Aléatoire
	Pourcentage de variables sélectionnées	10 %
Évaluation	Fonction à minimiser	RMSECV
	Type de validation croisée	partielle (10 itérations)
	Nombre maximum de facteurs PLS	20
Sélection	Méthode de sélection	"Roulette"
Recombinaison	Pourcentage de recombinaison	50%
Mutation	Pourcentage de mutation	0,5%
Réinsertion	Méthode de réinsertion	Élitiste
Critère de convergence	Pourcentage maximum de duplicatas	50%
	Nombre maximum de générations	150

utilisé. Ainsi, à chaque génération, 128 nouveaux individus vont être créés. La méthode par "roulette" est appliquée pour la sélection des 128 individus "parents". En utilisant cette méthode, tous les individus de la génération peuvent participer à la création de la nouvelle génération. La convergence de l'algorithme est ainsi ralentie ce qui permet d'éviter la convergence vers un optimum local. Les individus les plus adaptés ont néanmoins une probabilité plus forte d'être sélectionnés afin d'assurer l'adaptation de la population de génération en génération. Les nouveaux individus sont ensuite créés par croisement simple entre chromosomes des individus parents et une probabilité de mutations de 0,5% est utilisée. Enfin, une réinsertion élitiste est appliquée pour la création de la nouvelle génération qui se compose donc des 128 nouveaux individus et des 128 individus les plus adaptés de la génération précédente.

La convergence de l'optimisation est atteinte lorsque la population comporte 50% de

duplicatas ou lorsqu'un maximum de 150 générations est atteint.

4.2.1.2 Démarche d'exploitation des résultats

Pour chaque propriété, nous avons procédé à plusieurs optimisations par AG (Figure 4.10). Nous avons tout d'abord considéré uniquement l'optimisation des prétraitements. Les modèles obtenus par cette approche sont notés P-GAPLS. Nous avons ensuite procédé à l'optimisation simultanée des prétraitements et des variables à sélectionner. Ces modèles sont notés Co-GAPLS. Pour chacune de ces deux approches nous avons fait varier le nombre maximum de prétraitement p de 2 à 4. Enfin, pour chaque valeur de p, nous avons procédé à 5 exécutions de l'optimisation. En effet, l'optimisation d'un modèle sur une seule exécution n'est pas suffisant pour plusieurs raisons. Tout d'abord, le fait de procéder à plusieurs exécutions permet de réduire les risques de convergence vers un optimum local. De plus, lors de l'optimisation par AG, certaines des variables sélectionnées peuvent être dues à des corrélations aléatoires [56]. Ainsi, Leardi [67] propose de procéder à plusieurs exécutions de l'optimisation et de définir les variables du modèle final selon leur fréquence de sélection.

Ainsi, pour chaque jeu de paramètres (propriété, approche d'optimisation (P-GAPLS ou co-GAPLS), valeur de p), les cinq exécutions de la procédure fournissent 1280 solutions potentielles ($256 \times 5 = 1280$). Une démarche a donc été mise en œuvre pour sélectionner un modèle final à partir de ces solutions potentielles.

Tout d'abord, nous ne conservons que les 128 individus les plus adaptés pour chacune des cinq exécutions. Nous calculons ensuite la fréquence de sélection de chaque combinaison de prétraitements. Les individus qui admettent la combinaison de prétraitement la plus fréquemment sélectionnée sont conservés. Enfin, dans le cas de la co-optimisation, nous calculons la fréquence de sélection de chaque groupe de variables et éliminons ceux qui ont une fréquence de sélection inférieure à 20%. L'élimination de ces groupes de variables est effectuée dans le but de réduire le risque de présence de variables aléatoirement corrélées. Les prétraitements et les variables ainsi définis sont ensuite utilisés pour développer le modèle PLS final.

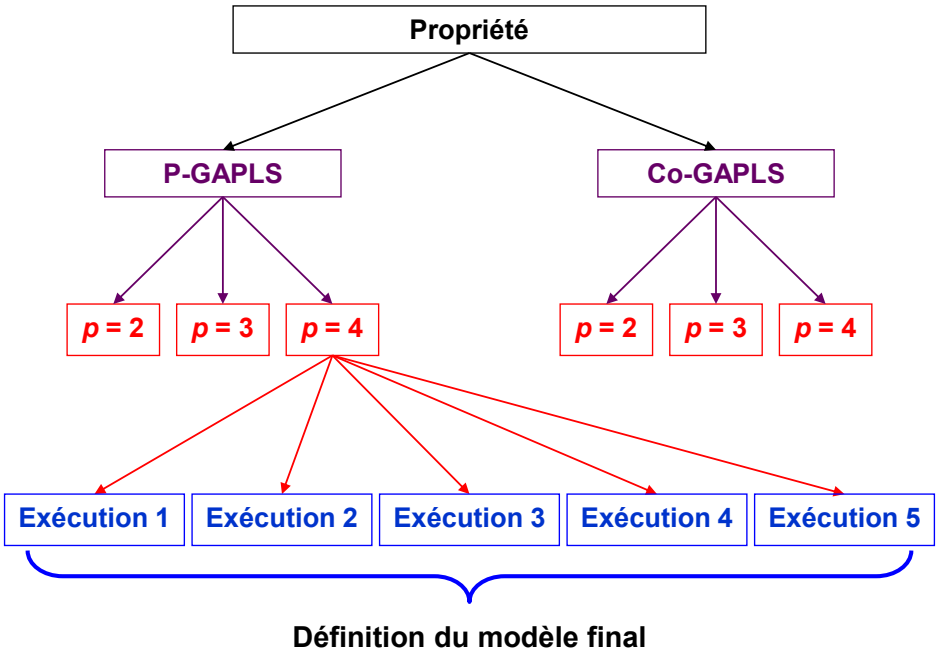


FIGURE 4.10 – Démarche de l'optimisation par AG pour chaque propriété. P-GAPLS : optimisation prétraitement seul, Co-GAPLS : optimisation simultanée du prétraitement et de la sélection de variables ; p : nombre maximum de prétraitement

Afin d'évaluer l'apport de la procédure mise en œuvre, des modèles dits "préliminaires" (noté PLS) ont été développés à partir du bilan des travaux effectués dans la littérature. En effet, dans la Partie 2.2, nous avons constaté que la majorité des modèles prédictifs des propriétés des produits lourds ont été développés à partir de la régression PLS, en appliquant la dérivée comme prétraitement et sur des domaines spectraux continus. Ainsi, la première étape de cette étude a consisté, pour chacune des propriétés considérées, à :

- optimiser les paramètres de l'algorithme de Savitzky-Golay pour le calcul de la dérivée (ordre de la dérivée, ordre du polynôme et la taille de la fenêtre spectrale),
- sélectionner le domaine spectral,
- fixer le nombre de facteurs PLS.

La procédure d'optimisation ainsi que le calcul de l'équation d'étalonnage des modèles PLS et AG finaux ont été réalisés sur le lot d'échantillons d'étalonnage, qui représente 75% des échantillons disponibles. Les échantillons d'étalonnage ont été choisis de manière aléatoire, les autres approches testées s'étant révélées inefficaces. En effet, des méthodes de sélection d'échantillons telles que les méthodes KENNARD et STONE et SPXY (Annexe

A.2) ont été appliquées sur les scores de l'ACP calculés sur les spectres PIR. Cependant, ces méthodes, qui sont basées sur le calcul de distances euclidiennes, ont amené à une séparation des échantillons en lots d'étalonnage et de validation non-optimale du fait de la présence de deux groupes d'échantillons. En effet, les échantillons de validation se regroupaient aux centres des deux groupes d'échantillons, ce qui se traduisait par une valeur de RMSEP jugée trop optimiste.

Afin de d'évaluer l'apport de la procédure d'optimisation par AG, les modèles préliminaires (PLS), P-GAPLS et Co-GAPLS sont comparés à l'aide du test "randomisation t-test". Ce test compare les modèles deux à deux sur les erreurs de prédiction des échantillons du lot de validation. Au cours de ce travail, un risque de 5% est utilisé et 10000 itérations sont réalisées pour le calcul de la valeur de *p-value* (Partie 3.5.1).

Pour finir, une analyse détaillée de chaque groupe de variables sélectionnées n'est pas envisageable du fait de la complexité des spectres PIR et de la composition chimique des produits lourds. Nous avons donc mis en œuvre une démarche pour aider à l'interprétation globale des modèles définis par les AG. Les principales bandes d'absorption des spectres PIR sont tout d'abord classées en trois groupes, selon le degré de saturation du groupement chimique auxquelles elles sont dues :

1. les bandes dues aux vibrations des liaisons dans les groupements chimiques saturés :
   - **4460-4000 $\mathbf{cm^{-1}}$** : combinaison entre les déformations des liaisons C-H et les élongations symétriques et asymétriques des liaisons C-H dans les groupements CH_2 et CH_3.
   - **6200-5400 $\mathbf{cm^{-1}}$** : première harmonique des élongations symétriques et asymétriques des liaisons C-H dans les groupements CH_2 et CH_3.
2. la bande due aux vibrations des liaisons de groupements insaturés :
   - **4700-4540 $\mathbf{cm^{-1}}$** : combinaison entre les élongations des doubles liaisons C=C et les élongations des liaisons =C-H à caractère aromatique.
3. la bande d'absorption due aux agrégats d'asphaltènes :
   - **9000-6000 $\mathbf{cm^{-1}}$** : l'absorption électronique des liaisons $\pi - \pi^*$ et $n - \pi^*$.

Le pourcentage de variables situées dans chaque groupe est ensuite calculé pour chaque modèle de prédiction. Une interprétation de ces pourcentages est ensuite effectuée par

comparaison entre les modèles des quatre propriétés considérées.

4.2.2 Résultats et discussion

Les résultats obtenus pour l'optimisation par AG du prétraitement et des variables à sélectionner sont présentés ici. Nous discuterons tout d'abord du choix du nombre maximum de prétraitements. Les modèles finaux de prédiction obtenus dans le cadre de l'optimisation du prétraitement (P-GAPLS) et de la co-optimisation (Co-GAPLS) sont ensuite décrits. Enfin, une interprétation des variables sélectionnées est effectuée.

4.2.2.1 Nombre maximum de prétraitements

Pour chaque propriété et pour chaque approche d'optimisation (P-GAPLS et Co-GAPLS), nous avons réalisé des calculs en faisant varier le nombre maximum p de prétraitements de 2 à 4. En effet, au moins deux prétraitements sont nécessaires car nous avons décidé de permettre à l'algorithme d'optimiser le choix entre plusieurs types de centrage et de mises à l'échelle (Tableau 3.2, Partie 3.3.3).

Nous avons observé que, lorsque $p = 2$, les méthodes sélectionnées pour les modèles finaux sont toujours un centrage ou une mise à l'échelle combiné à une correction de spectres. Certains individus présentaient cependant des combinaisons de prétraitements paradoxales lorsque le nombre maximum de prétraitements p est égal à 3 ou 4. En effet, deux centrages par la moyenne ou deux dérivées $1^{\text{ère}}$ pouvaient être sélectionnés dans la même solution. De ce fait, ces solutions sont aberrantes. Nous présenterons alors que les optimisations pour lesquelles $p = 2$.

4.2.2.2 Performances des modèles optimisés par AG

Cette partie a pour but de présenter les modèles développés par PLS, P-GAPLS et Co-GAPLS pour la détermination des teneurs en saturés, en aromatiques, en résines et en asphaltènes C7. Le Tableau 4.3 indique, pour chaque modèle, les prétraitement appliqués, les variables (ou le domaine spectral) retenues, le nombre de facteurs PLS et les performances obtenues en termes d'erreur de prédiction. Les résultats de la comparaison des modèles obtenus par le test "randomisation t-test" sont également résumés dans le

Tableau 4.3. Pour rappel, deux modèles sont significativement différents si la valeur de "p-value" est inférieure ou égale à 0,05.

Lors de l'optimisation des paramètres des modèles préliminaires (PLS), les erreurs de RMSEP les plus faibles ont été obtenues en appliquant aux spectres une dérivée 1$^{\text{ère}}$ calculée avec un polynôme d'ordre 2 et une taille de fenêtre de 15 points. De plus, les modèles de prédiction des teneurs en saturés, en aromatiques et en résines ont été développés sur le domaine spectral 7000-4000 cm^{-1} qui permet de pallier les problèmes liés à l'absorption électronique. Pour la prédiction en asphaltènes, l'équation d'étalonnage est calculée sur le domaine 9000-4000 cm^{-1}. Le nombre de facteurs PLS fixé pour chacun des modèles est indiqué dans le Tableau 4.3.

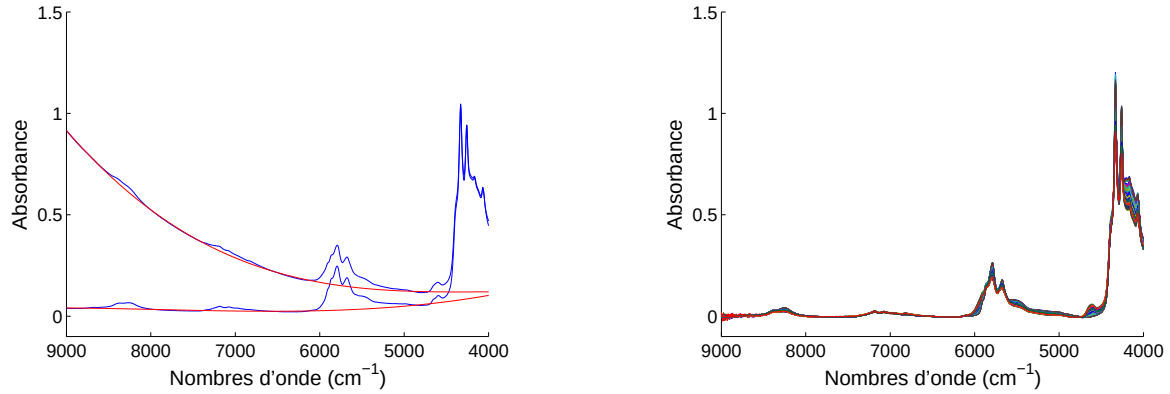
Avant de présenter en détail les résultats obtenus lors de l'optimisation par AG, nous voulons tout d'abord noter que cette approche a été testée sur le domaine spectral 9000-4000 cm^{-1} pour les modèles de prédiction des teneurs en saturés, en aromatiques et en résines. Les résultats obtenus n'étaient néanmoins pas concluants. Le domaine spectral 7000-4000 cm^{-1} a donc ici aussi été utilisé pour les optimisations par AG de ces modèles.

Pour la prédiction de la teneur en saturés, le modèle final retenu lors de l'optimisation P-GAPLS est développé sur 8 facteurs PLS. Les performances obtenues en termes d'erreurs de RMSEC et RMSEP sont respectivement de 1,31 et 1,51 %(m/m). La combinaison de prétraitements sélectionnée est la correction *Weighted Least Square Baseline* (WLSB) avec un polynôme d'ordre 3 et un centrage par la moyenne. Le principe de la correction WLSB (Annexe A.1.2.2) est d'ajuster la ligne de base, ici à l'aide d'un polynôme d'ordre 3, et de la retirer au spectre (Figure 4.11a). La Figure 4.11b illustre les spectres PIR de la base d'échantillons sur le domaine 9000-4000 cm^{-1} corrigés par ce prétraitement. Nous constatons que cette méthode permet de corriger très efficacement les variations de la ligne de base engendrées par l'absorption électronique.

Tableau 4.3 – Présentation et comparaison des modèles PLS, P-GAPLS et Co-GAPLS pour la prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7

Propriété	**Approche**	**Nombre de variables (Domaine spectral (cm^{-1}))**	**Prétraitements**	**Nombre de facteurs PLS**	**RMSEC %(m/m)**	**RMSEP %(m/m)**	***p*-value**	
							vs PLS[d]	**vs P-GAPLS[e]**
Saturés	PLS	1557 (7000-4000)	Dérivée 1[ère] + CM[a]	7	1,75	2,09	-	-
	P-GAPLS	1557 (7000-4000)	WLSB[b] (3[c]) + CM[a]	8	1,31	1,51	$\leq 0,05$	-
	Co-GAPLS	460	WLSB[b] (3[c]) + CM[a]	9	1,38	1,82	0,14	0,07
Aromatiques	PLS	1557 (7000-4000)	Dérivée 1[ère] + CM[a]	6	1,86	2,06	-	-
	P-GAPLS	1557 (7000-4000)	WLSB[b] (3[c]) + CM[a]	9	1,17	1,68	0,07	-
	Co-GAPLS	520	CM[a] + Detrend (2[c])	10	0,99	1,59	$\leq 0,05$	0,63
Résines	PLS	1557 (7000-4000)	Dérivée 1[ère] + CM[a]	8	1,24	0,82	-	-
	P-GAPLS	1557 (7000-4000)	WLSB[b] (3[c]) + CM[a]	9	0,88	0,80	0,45	-
	Co-GAPLS	500	WLSB[b] (1[c]) + Autoscale	9	0,86	0,77	0,40	0,62
Asphaltènes C7	PLS	2604 (9000-4000)	Dérivée 1[ère] + CM[a]	3	1,41	1,26	-	-
	P-GAPLS	1604 (9000-4000)	Dérivée 1[ère] + CM[a]	3	1,41	1,26	1	-
	Co-GAPLS	600	Detrend (2[c]) + CM[a]	8	0,63	1,28	0,24	0,24

[a] : CM = Centrage par la moyenne
[b] : WLSB = Weighted Least Square Baseline
[c] : Ordre du polynôme
[d] : Comparaison des modèles P-GAPLS et Co-GAPLS avec le modèle PLS
[e] : Comparaison du modèle P-GAPLS avec le modèle Co-GAPLS



(a) Exemple d'ajustement de la ligne de base par le prétraitement WLSB avec un polynôme d'ordre 3

(b) Spectres de la base corrigés par le prétraitement WLSB avec un polynôme d'ordre 3

FIGURE 4.11 – Le prétraitement *Weighted Least Square Baseline* (WLSB)

Les résultats obtenus par le test de comparaison indique que le modèle P-GAPLS est significativement plus performant que le modèle PLS dont l'erreur de RMSEP est égal à 2,09 %(m/m). En effet, une valeur de *p-value* inférieure à 0,05 a été obtenue. Nous pouvons donc conclure que, par rapport à l'application d'une dérivée 1$^{\text{ère}}$, le prétraitement WLSB permet d'améliorer significativement le pouvoir prédictif du modèle de détermination de la teneur en saturés. Ce prétraitement a également été sélectionné lors des optimisations P-GAPLS effectuées dans le cadre de la prédiction des teneurs en aromatiques et en résines.

Le prétraitement WLSB avec un polynôme d'ordre 3 a aussi été sélectionné lors de la co-optimisation (Co-GAPLS) du modèle de prédiction des teneurs en saturés. Pour cette approche, 460 variables ont été retenues et 9 facteurs PLS ont été utilisés. Cependant, la valeur de RMSEP obtenue de 1,82 %(m/m) n'est pas significativement différente de celle obtenue par les modèles PLS et P-GAPLS.

Les erreurs de RMSEC et de RMSEP du modèle P-GAPLS pour la prédiction des aromatiques, développé avec 9 facteurs PLS, sont respectivement de 1,17 et 1,68 %(m/m). Le test de comparaison n'a pas mis en évidence de différence significative entre cette valeur de RMSEP et celle obtenue par le modèle PLS, qui est de 2,06 %(m/m) (p-value = 0,07). En revanche, le pouvoir prédictif du modèle Co-GAPLS, pour lequel l'erreur de RMSEP

obtenue est de 1,59 %(m/m), est considéré significativement meilleur que celui du modèle PLS (p-value $\leq 0{,}05$).

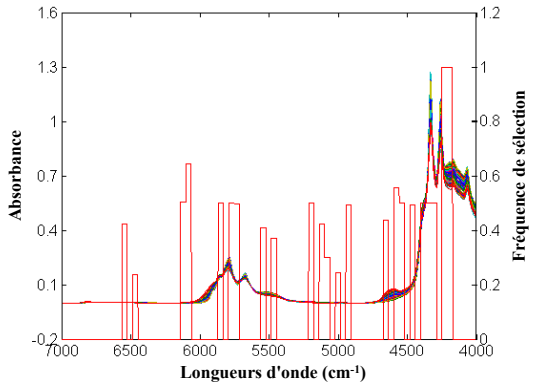
Pour la détermination de la teneur en résines, les valeurs obtenues pour l'erreur RMSEP sont comparables pour les trois approches : 0,82 %(m/m) pour le modèle PLS, 0,80 %(m/m) pour l'approche P-GAPLS et 0,77 %(m/m) pour le modèle défini par la procédure Co-GAPLS. Par conséquent, aucune différence significative entre ces trois modèles n'a pu être démontrée.

Enfin, la dérivée 1$^{\text{ère}}$ et le centrage par la moyenne ont été sélectionnés pour la prédiction des teneurs en asphaltènes C7 lors de l'optimisation du prétraitement (P-GAPLS). Par conséquent, nous pouvons noter que les modèles PLS et P-GAPLS sont identiques. Nous pouvons donc en déduire que ce prétraitement est le meilleur compromis pour réduire la variabilité de l'absorption électronique sans l'éliminer lorsque le domaine spectral continu est considéré. Ces modèles ont été développés avec 3 facteurs PLS. Une valeur de 1.26 %(m/m) a été obtenue pour l'erreur RMSEP. Celle obtenue par le biais du modèle Co-GAPLS est du même ordre de grandeur (*i.e.* 1,28 %(m/m)). Cependant, ce modèle est moins parcimonieux car 8 facteurs PLS sont nécessaires.

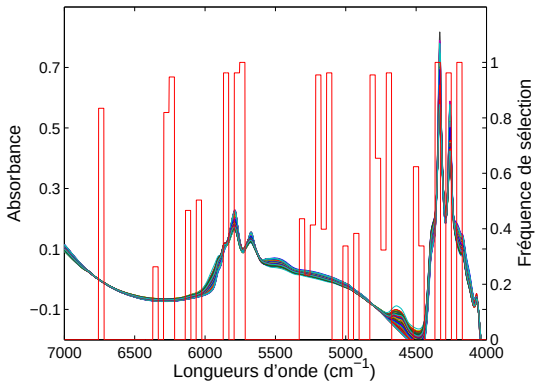
4.2.2.3 Interprétation de la sélection de variables

Nous interprétons dans cette partie les variables sélectionnées lors de la co-optimisation par AG (Co-GAPLS). En effet, bien que cette optimisation n'ait pas systématiquement apporté d'amélioration significative des capacités de prédiction, le potentiel d'interprétation lié à la sélection de variables peut être considéré comme un avantage notable.

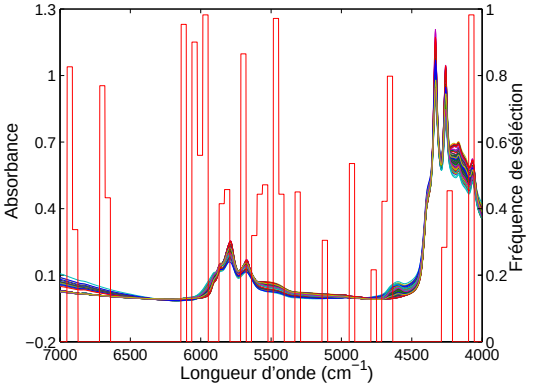
La Figure 4.12 illustre les variables sélectionnées et la fréquence de sélection associée pour chaque propriété. Les spectres représentés sur cette figure sont issus de l'application du prétraitement déterminé lors de la procédure de co-optimisation par AG. Il faut noter que le prétraitement correspondant au centrage ou à la mise à l'échelle des données spectrales n'a pas été appliqué ici afin de faciliter l'interprétation.



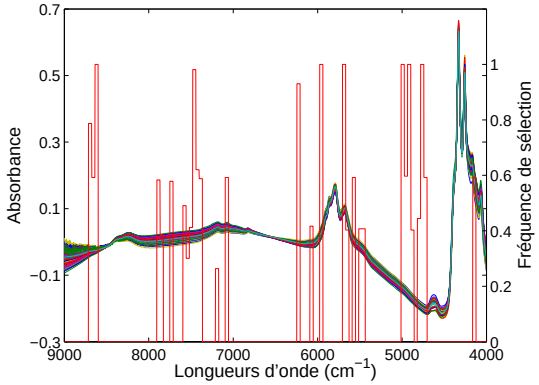
(a) Modèle Co-GAPLS pour la détermination de la teneur en saturés ; 460 variables sélectionnées ; spectres prétraités par la méthode WLSB à l'ordre 3



(b) Modèle Co-GAPLS pour la détermination de la teneur en aromatiques ; 520 variables sélectionnées ; spectres prétraités par la méthode *Detrend* à l'ordre 2



(c) Modèle Co-GAPLS pour la détermination de la teneur en résines ; 500 variables sélectionnées ; spectres prétraités par la méthode WLSB à l'ordre 1



(d) Modèle Co-GAPLS pour la détermination de la teneur en asphaltènes C7 ; 600 variables sélectionnées ; spectres prétraités par la méthode *Detrend* à l'ordre 2

FIGURE 4.12 – Interprétation des variables sélectionnées dans le cadre de la co-optimisation par AG ; Le prétraitement sélectionné lors de co-optimisation par AG est appliqué ; les blocs de variables sélectionnés sont représentés par les barres rouges et leur intensité correspond à la fréquence de sélection

Du fait de la complexité des spectres PIR et de la composition chimique des produits lourds, l'interprétation détaillée de toutes les variables sélectionnées n'est pas envisageable. Ainsi, nous avons mis en œuvre une démarche qui consiste à répartir les bandes d'absorption présentes dans les spectres PIR au sein de quatre groupes différents selon des caractéristiques chimiques identifiables (Partie 4.2.1.2) :

– **groupe 1** : bandes dues aux vibrations de liaison dans les groupements saturés

– **groupe 2** : bandes dues aux vibrations de liaison dans les groupements insaturés
– **groupe 3** : zone spectrale de la courbure de la ligne de base (9000-6000 cm^{-1})
– **groupe 4** : zones spectrales non attribuées.

Les blocs de variables sélectionnées sont par la suite attribués à chacun de ces groupes. Enfin, une interprétation globale des modèles Co-GAPLS de prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7 est réalisée en comparant la répartition des blocs de variables dans chacun de ces groupes. Le Tableau 4.4 résume la répartition des blocs de variables sélectionnées.

Le modèle Co-GAPLS de prédiction des saturés est développé en utilisant 460 variables (23 blocs). Les prétraitements appliqués sont la correction WLSB avec un polynôme de degré 3 et un centrage de la moyenne. Comme nous l'avons mentionné précédemment, ce prétraitement permet une correction très efficace de la ligne de base. La moitié des variables sélectionnées (12 blocs) sont localisées sur les bandes dues aux vibrations de liaison de groupements saturés (groupe 1). De plus, 2 blocs de variables sont situés dans la région due aux vibrations de groupements insaturés (groupe 2). Les coefficients PLS sont négatifs dans cette région. Nous pouvons donc supposer qu'il s'agit d'une corrélation inverse. Enfin, 9 blocs de variables sont situés dans des régions qui n'ont pas pu être attribuées.

Pour le modèle Co-GAPLS de prédiction des aromatiques, 520 variables ont été sélectionnées (26 blocs). Nous pouvons noter que le pourcentage de variables situées sur les bandes attribuées aux groupements chimiques saturés (groupe 1) diminue par rapport au modèle de prédiction de la teneur en saturés. En revanche, le nombre de blocs situés sur les bandes causées par les groupements chimiques insaturés augmente (groupe 2). Enfin, un nombre important de variables ne sont pas attribuées (55%). La Figure 4.12b illustre les spectres issus de l'application de la correction *Detrend* avec un polynôme d'ordre 2. Ce prétraitement réduit fortement les variabilités dues à la courbure de la ligne de base. Cependant, contrairement aux spectres prétraités par la méthode WLSB à l'ordre 3, on observe une forte variabilité dans les zones qui n'ont pas été attribuées (7000-6500 cm^{-1} et 5400-4540 cm^{-1}). Ainsi, nous pouvons supposer la présence d'informations dans ces zones ce qui justifierait le pourcentage élevé de blocs situés sur ces domaines spectraux.

Tableau 4.4 – Localisation des blocs de variables sélectionnés

	Saturés	**Aromatiques**	**Résines**	**Asphaltènes C7**
Nombre de variables (Nombre de blocs)	460 (23)	520 (26)	500 (25)	600 (30)
Nombre de blocs situés sur le groupe 1[a]	12 (52%)	8 (30%)	12 (44%)	6 (22%)
Nombre de blocs situés sur le groupe 2[b]	2 (9%)	4 (15%)	2 (8%)	3 (3%)
Nombre de blocs situés sur le groupe 3[c]	0	0	4 (16%)	14 (45%)
Nombre de blocs situés sur le groupe 4[d]	9 (39%)	14 (55%)	8 (32%)	7 (23%)

[a] : Groupe 1 : bandes dues aux groupements saturés
[b] : Groupe 2 : bandes dues aux groupements insaturés
[c] : Groupe 3 : zone spectrale de la courbure de la ligne de base
[d] : Groupe 4 : zones spectrales non interprétées

Le prétraitement sélectionné pour le modèle de prédiction de la teneur en résines est la méthode WLSB à l'ordre 1. La Figure 4.12c montre que les variations dues à la courbure de la ligne de base ne sont pas entièrement corrigées. De plus, parmi les 25 blocs sélectionnés pour ce modèle (500 variables), 4 sont localisés sur ce domaine. La répartition des autres blocs est la suivante : 12 blocs sur le groupe 1, 2 blocs sur le groupe 2 et 8 blocs sur le groupe 4.

Le modèle Co-GAPLS de prédiction de la teneur en asphaltènes est développé sur 600 variables (30 blocs) avec la correction *Detrend* (polynôme d'ordre 2). Nous pouvons remarquer sur la Figure 4.12d que, du fait du domaine spectral considéré pour ce modèle (9000-4000 cm^{-1}), les spectres obtenus sont très différents de ceux obtenus pour le modèle

de prédiction des teneurs en aromatiques (domaine spectral : 7000-4000 cm^{-1}). De plus, en appliquant cette méthode, la variabilité de la courbure de la ligne de base, bien que fortement réduite, n'est pas totalement éliminée. D'ailleurs, 45% des variables sélectionnées se trouvent dans cette zone (groupe 3). Nous pouvons également noter que le nombre de blocs situés sur les bandes correspondant aux vibrations des groupements saturés et insaturés est faible et, respectivement, de 6 et de 3.

Après avoir attribué les variables sélectionnées dans les différents groupes, la répartition des blocs est examinée afin de procéder à une interprétation globale. Il faut rappeler que, lors du fractionnement SARA, les molécules sont séparées en fonction de leur polarité et donc en partie de leur degré de saturation.

Le Tableau 4.4 indique que le pourcentage de variables situées sur les bandes dues aux vibrations de liaisons dans les groupements saturés (groupe 1) diminuent globalement de la fraction des saturés à la fraction des asphaltènes C7. Ce résultat apparaît en accord avec la composition chimique des fractions SARA car le degré de saturation augmente lorsque l'on passe de la fraction des saturés à la fraction des asphaltènes. De plus, le pourcentage maximum de variables sélectionnées dans la zone des bandes dues aux vibrations de liaisons dans les groupements insaturés (groupe 2) est obtenu pour la fraction aromatique. Ce résultat peut paraître en contradiction avec la composition chimique des fractions. En effet, il aurait été logique que ce nombre augmente pour les fractions résines et asphaltènes C7. La sélection d'un grand nombre de variables sur le domaine spectral de l'absorption électronique (groupe 3) pour ces deux propriétés peut néanmoins expliquer ce constat. Nous avons vu que les prétraitements sélectionnés pour les modèles de prédiction des résines et des asphaltènes C7 ne corrigeaient pas totalement les variations de la courbure de la ligne de base. Pour le modèle de détermination de la teneur en asphaltènes C7, il est logique qu'un grand nombre de variables sélectionnées se situent sur ce domaine spectral. Pour le modèle de prédiction de la teneur en résines, 4 blocs ont également été sélectionnés dans cette zone. Nous avons évoqué dans la Partie 4.1.1 que la teneur en résines est extrêmement corrélée à la présence (ou l'absence) d'asphaltènes. De plus, il a été démontré que les résines jouent un rôle très important pour la stabilité des asphaltènes [96]. L'absorption électronique peut alors s'avérer être une information utile

pour la détermination de la teneur en résines et, peut justifier le prétraitement fixé et les variables sélectionnées sur ce domaine.

Cette interprétation indique que les variables sélectionnées sont globalement en adéquation avec la composition chimique des fractions SARA. Pour résumer, le pourcentage de variables situées sur les bandes dues aux groupements chimiques saturés diminue de la fraction des saturés à de la fraction des des asphaltènes. De plus, le nombre de blocs localisés sur les bandes dues aux groupements insaturés est maximum pour la fraction des aromatiques. Enfin, le nombre de variables sélectionnées dans la zone de l'absorption électronique est très élevé pour le modèle de prédiction des asphaltènes C7. De plus, la présence de variables dans cette zone pour le modèle des résines peut-être interprétée.

Un grand nombre de variables est néanmoins situé sur des zones que l'on ne peut pas attribuer à la vibration d'un groupement particulier. Ceci est, d'une part, dû au fait que de très faibles bandes peuvent être observées dans les spectres PIR. D'autre part, certains groupes de variables ont des coefficients proches de zéro. Il est alors légitime de penser que l'influence de ces groupes de variables est faible et qu'il rende l'interprétation difficile. Une méthode d'évaluation de l'apport de chaque variable à la description de la propriété à modéliser pourrait alors être utilisée afin de s'assurer de l'apport de chaque groupe de variables sélectionnées aux modèles.

Leardi *et al.* [68] ont introduit une nouvelle approche, la méthode *backward interval partial least squares* (BiPLS). Cette approche a pour but, en amont de l'optimisation de la sélection de variables par AG, d'éliminer les zones spectrales les moins pertinentes. Ainsi, l'espace des solutions potentielles est réduit à un nombre de variables que les AG peuvent traiter plus facilement. Le principe de la BiPLS est de grouper les variables en N blocs et de réaliser une PLS en considérant chacune des combinaisons de N-1 blocs. Le bloc qui n'est pas présent dans la combinaison des N-1 blocs ayant amenée à l'erreur de validation croisée la plus faible est éliminé.

Cette approche ne peut néanmoins pas être mise en œuvre dans le cadre de notre procédure d'optimisation. En effet, l'application de cette méthode en amont nécessite de fixer un prétraitement, ce qui va à l'encontre de la philosophie de l'optimisation simultanée du prétraitement et de la sélection de variables. Pour une implémentation future de cette

procédure de co-optimisation par AG, nous proposons la BiPLS d'appliquer au modèle final pour s'assurer de l'apport de chaque variables sélectionnées à la description de la propriété. L'interprétation des modèles serait ainsi plus juste. Une autre solution est d'introduire la BiPLS à des intervalles réguliers de générations au cours de l'optimisation afin de permettre une convergence plus rapide de la procédure.

4.2.3 Bilan

Cette partie avait pour but de présenter l'étude réalisée pour l'optimisation des modèles développés à partir des spectres PIR. En effet, nous avons constaté la présence de deux groupes d'échantillons lors de la description par ACP des spectres PIR. De plus, nous avons mentionné que les fortes variations engendrées par l'absorption électronique devaient probablement être réduites même si elles étaient une information pertinente dans le cadre de la prédiction des teneurs en asphaltènes C7. Nous avons évoqué l'importance pour ce modèle de l'interaction entre le prétraitement et les variables à sélectionner. Nous avons donc évalué l'apport de l'optimisation du prétraitement et de la sélection de variables par AG.

Premièrement, cette procédure d'optimisation apporte des réponses concernant le choix du prétraitement. La dérivée 1ère a été sélectionnée pour le modèle de prédiction de la teneur en asphaltènes C7 lorsque nous avons considéré uniquement l'optimisation du prétraitement. Cette correction apparaît donc comme la plus adaptée pour réduire la variabilité de la courbure de la ligne de base tout en conservant l'information qu'elle contient. L'optimisation du prétraitement par AG a également permis de mettre en évidence que l'application du prétraitement "*Weighted Least Square Baseline*" (WLSB) avec un polynôme d'ordre 3 permet une correction très efficace de la courbure de la ligne de base présente dans les spectres PIR. Ce prétraitement a été sélectionné pour les modèles de prédiction des teneurs en saturés, en aromatiques et en résines. De plus, pour la détermination de la teneur en saturés, il a été démontré que l'application de la méthode WLSB aux spectres PIR permet une amélioration du pouvoir prédictif par rapport au modèle développé à partir de la dérivée 1ère. La valeur de RMSEP est également plus faible pour les modèles de prédiction des teneurs en aromatiques et en résines développés

avec la méthode WLSB même si aucune différence significative n'a pu être démontrée. Nous pouvons donc conclure que l'application de la méthode WLSB permet globalement une meilleure correction des spectres PIR par rapport à la dérivée 1$^{\text{ère}}$ lorsque l'on cherche à éliminer la courbure de la ligne de base. La procédure mise en œuvre a donc permis de fixer les prétraitements qui conduisent aux meilleurs pouvoir prédictifs.

Deuxièmement, l'approche proposée pour l'optimisation du prétraitement et de la sélection de variables permet, dans certains cas, une amélioration de la prédiction ainsi qu'une interprétation de l'information chimique. En effet, une amélioration significative du pouvoir prédictif pour la détermination de la teneur en aromatiques a été démontrée par le modèle Co-GAPLS comparé au modèle PLS. Pour les modèles Co-GAPLS des autres propriétés, les erreurs de RMSEP obtenues sont du même ordre de grandeur que celles obtenues avec les autres approches. De plus, nous avons constaté que les variables sélectionnées pour les modèles de prédictions des teneurs en SARA sont globalement en accord avec la composition chimique de ces fractions.

Enfin, nous avons montré les limites de la procédure mise en œuvre. Des implémentations ont donc été proposées afin d'améliorer la cette approche d'optimisation. Nous avons évoqué que des combinaisons paradoxales apparaissaient lorsque le nombre maximum de prétraitements était égal à 3 ou 4. Ceci est du au fait que les prétraitements étaient choisis aléatoirement. Pour pallier ce problème, il faudrait envisager d'introduire des restrictions sur le choix des prétraitements. Pour cela, nous proposons de grouper les différents prétraitements par types et d'interdire la combinaison de prétraitements appartenant au même groupe. Les différents groupes sont les suivants : filtrage, normalisation, dérivation, correction de la ligne de base et centrage/mise à l'échelle (Tableau 3.2, Partie 3.3.3). Il serait aussi envisageable de forcer l'algorithme à appliquer un centrage ou une mise à l'échelle. Nous avons également observé qu'un grand nombre de variables sélectionnées étaient situées dans des zones non attribuées. La sélection de ces variables n'est pas à remettre en question car de très faibles bandes d'absorption peuvent être présentes dans les spectres PIR. De plus, les coefficients de certains blocs étaient très faibles dans ces régions. Nous proposons tout de même de mettre en place la méthode BiPLS pour évaluer l'apport pour le modèle de chaque variable sélectionnée.

4.3 Comparaison et fusion des spectroscopies MIR et PIR

Nous avons évoqué que, dans la littérature, les modèles développés à partir des spectroscopies MIR et PIR conduisent à des performances différentes bien que la majeure partie de l'information contenue dans ces deux domaines soient dues aux vibrations des liaisons dans les mêmes groupements chimiques (Partie 2.1.3). Nous avons également mentionné la difficulté de tirer des conclusions quant au potentiel de ces deux techniques spectroscopiques à partir des études comparatives. En effet, les résultats obtenus dépendent fortement de l'instrumentation utilisée, des produits analysés, des gammes analytiques des propriétés considérées et des techniques chimiométriques mise en œuvre. Nous proposons alors d'évaluer les pouvoirs prédictifs des modèles développés séparément à partir des spectroscopies MIR et PIR.

L'interprétation des spectres a révélé que ces deux spectroscopies contiennent de l'information spécifique. D'une part, une bande d'absorption électronique a été observée dans le domaine du PIR. D'autre part, nous avons constaté la présence des vibrations fondamentales des liaisons C-H hors du plan en spectroscopie MIR. Nous pouvons ainsi supposer que la fusion de ces informations permettra de mieux décrire les propriétés considérées. Dans la Partie 3.4, nous avons mentionné les limites d'un développement de modèle PLS de fusion de données à partir de la concaténation des bases spectrales. Nous avons donc décidé d'évaluer deux méthodes *multiblock*, la *multiblock* PLS (MB-PLS) et la *serial* PLS (S-PLS). La philosophie de ces méthodes est de séparer les variables en différents blocs. Ici, deux blocs sont définis, un bloc pour les spectres MIR et un pour les spectres PIR. La MB-PLS traite ensuite les blocs en parallèle tandis que la S-PLS les traite en série. Le principal avantage de ces méthodes est qu'elles permettent de définir la contribution de chaque spectroscopie à la description de la propriété considérée. L'apport de l'exploitation simultanée des spectroscopies MIR et PIR par les méthodes MB-PLS et S-PLS pour la prédiction de propriétés des produits lourds sera donc évalué dans cette partie. Tout d'abord, nous décrirons la démarche mise en œuvre pour cette étude. Ensuite, nous comparons les performances des modèles et une interprétation des résultats sera réalisée.

4.3.1 Démarche

Afin de permettre une comparaison des différentes approches, la séparation des échantillons en lots d'étalonnage et de validation doit être identique. Nous précisons tout d'abord que la séparation considérée lors des optimisations par AG n'a pas été utilisée pour cette étude. Lors du développement des modèles MIR, de nouveaux échantillons aberrants sont apparus. La répartition des échantillons dans les lots d'étalonnage et de validation n'était, par conséquent, plus équilibrée. De plus, certains échantillons se situaient à l'extrémité du nuage de points des scores de l'ACP calculée sur les spectres MIR. Afin de ne privilégier aucune des deux spectroscopies, une séparation uniquement basée sur la propriété a donc été employée. Pour ce faire, les valeurs de propriétés ont été classées par ordre croissant. Les échantillons du lot de validation ont ensuite été sélectionnés en tirant un échantillon sur quatre. Des précautions ont été prises afin de ne pas sélectionner les valeurs extrêmes de la gamme analytique. Cette procédure a été mise en œuvre car les méthodes de sélection basées sur les distances euclidiennes entre les spectres ne sont pas adaptées à cette étude. En effet, lorsque ces méthodes ont été appliquées à l'une de ces deux bases spectrales, il en résulte que la sélection n'est pas optimale pour l'autre base. Ceci se traduit notamment par la présence d'échantillons aux extrémités du nuage de points des scores de l'ACP.

Pour les modèles développés à partir des spectres PIR, nous avons opté pour les paramètres obtenus lors de l'optimisation du choix du prétraitement par AG. Les modèles de prédiction des teneurs en saturés, en aromatiques et en résines ont donc été développés sur le domaine 7000-4000 cm^{-1} en appliquant la méthode WLSB avec un polynôme d'ordre 3. Pour la prédiction des teneurs en asphaltènes C7, l'équation d'étalonnage a été établie à partir de la dérivée $1^{ère}$ des spectres sur le domaine 9000-4000 cm^{-1}. Les équations d'étalonnage calculées à partir des spectres MIR ont été développées en appliquant la dérivée $1^{ère}$ en considérant les domaines spectraux 3200-2700 cm^{-1} et 1700-650 cm^{-1}.

Pour chaque propriété, des modèles PLS ont été calculés en considérant séparément les spectres MIR et PIR. Les régressions MB-PLS et la S-PLS ont ensuite été utilisées pour la fusion des données spectrales. Pour fixer le nombre de facteurs, la validation croisée partielle a été employée. Pour chaque modèle, la RMSECV a donc été calculée

en fonction du nombre de facteurs a ($a = 1, \ldots, 15$). Nous avons évoqué dans la Partie 3.4.2 que l'ordre des blocs est important en S-PLS et que le nombre de facteurs de chaque bloc doit être fixé indépendamment. Les valeurs de RMSECV ont donc été calculées en effectuant toutes les combinaisons $a_{a1,a2}$ (avec a1 et a2 le nombre de facteurs des blocs 1 et 2, a1 et a2 = 1, ..., 15) et pour chaque combinaison d'ordre des blocs. En MB-PLS, le problème ne se pose pas car les blocs sont traités en parallèle. Le nombre de facteurs est donc commun aux deux blocs. Le test "randomisation t-test" a été appliqué pour comparer la qualité des prédictions obtenues par ces différentes approches. Comme évoqué dans la Partie 3.5.1, ce test compare les modèles deux à deux. Il faut également noter que nous avons calculé la valeur de *p-value* à partir de 10000 itérations.

4.3.2 Performances des modèles

Le Tableau 4.5 présente une synthèse des résultats obtenus par les différentes techniques de régression : le nombre de facteurs considérés, l'erreur de RMSEP et les valeurs de *p-value* obtenues. Pour chaque propriété, nous présenterons tout d'abord les résultats obtenus dans le cadre de la comparaison des modèles PLS développés séparément à partir des spectroscopies MIR et PIR. Nous évaluerons ensuite l'apport des méthodes d'exploitation simultanée des deux bases spectrales pour la prédiction des valeurs de ces propriétés. Nous voulons signifier que, pour les modèles PLS, nous nous permettrons un abus de langage en parlant de modèles MIR et PIR pour désigner les modèles PLS développés à partir de ces techniques spectroscopiques. En effet, ces dénominations permettent d'alléger fortement la discussion.

Pour la prédiction des teneurs en résines, les modèles MIR et PIR ont été développés avec respectivement 10 et 9 facteurs PLS. Les valeurs de RMSEP obtenues pour ces deux modèles sont respectivement de 1,01 %(m/m) et 1,10 %(m/m). Il a été établi que les performances des modèles MIR et PIR sont équivalentes.

Tableau 4.5 – Présentation et comparaison des modèles PLS, MB-PLS et S-PLS pour la prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7

Propriété	**Spectroscopie**	**Régression**	**Nombre de facteurs PLS**	**RMSEP**	***p*-value**		
				%(m/m)	**vs. PLS (MIR)**[b]	**vs. PLS (PIR)**[c]	**vs MB-PLS**[d]
Saturés	MIR	PLS	8	2,65	-	-	-
	PIR	PLS	5	1,68	$\leq 0,05$	-	-
	MIR + PIR	MB-PLS		1,79	$\leq 0,05$	0,17	-
	PIR + MIR [a]	S-PLS	8 + 5	1,87	$\leq 0,05$	$\leq 0,05$	0,18
Aromatiques	MIR	PLS	12	2,48	-	-	-
	PIR	PLS	9	1,47	$\leq 0,05$	-	-
	MIR + PIR	MB-PLS	9	2,02	$\leq 0,05$	$\leq 0,05$	-
	PIR + MIR[a]	S-PLS	15 + 4	1,24	$\leq 0,05$	0,19	$\leq 0,05$
Résines	MIR	PLS	10	1,01	-	-	-
	PIR	PLS	9	1,10	0,33	-	-
	MIR + PIR	MB-PLS	9	0,75	$\leq 0,05$	$\leq 0,05$	-
	MIR + PIR[a]	S-PLS	1 + 11	0,94	0,36	$\leq 0,05$	0,12
Asphaltènes C7	MIR	PLS	3	1,44	-	-	-
	PIR	PLS	3	1,05	0,22	-	-
	MIR + PIR	MB-PLS	4	0,97	0,065	0,18	-
	MIR + PIR[a]	S-PLS	3 + 3	0,94	0,095	0,15	0,37

[a] : Ordre des blocs en S-PLS
[b] : Comparaison des modèles PIR, MB-PLS et S-PLS au modèle modèle MIR
[c] : Comparaison des modèles MB-PLS et S-PLS au modèle modèle PIR
[d] : Comparaison des modèles MB-PLS et S-PLS

En effet, la valeur de *p-value* est de 0,33. Le modèle MB-PLS a été calculée avec 9 facteurs et l'erreur de RMSEP est de 0,75 %(m/m). Les valeurs de *p-value* obtenues lors de la comparaison de ce modèle MB-PLS avec les modèles MIR et PIR sont inférieures à 0,05. Nous pouvons en déduire que l'exploitation simultanée des spectroscopies MIR et PIR par la MB-PLS permet une amélioration significative des performances du modèle de prédiction des teneurs en résines. Enfin, nous pouvons noter que le modèle S-PLS a été développé en considérant les spectres MIR en premier bloc. La valeur de RMSEP de ce modèle (0,94 %(m/m)) est également inférieure à celles obtenues par les modèles MIR et PIR. Il a été démontré que les performances du modèle S-PLS sont meilleures que celles du modèle PIR. Aucune différence significative avec le modèle MIR n'a néanmoins pu être démontrée.

Les modèles MIR et PIR sont également équivalents pour la détermination des teneurs en asphaltènes C7 bien que les valeurs de RMSEP obtenues soient de 1,44 %(m/m) et 1,05 %(m/m) respectivement. En effet, il a été constaté que la forte valeur de RMSEP du modèle MIR est principalement due à deux échantillons mal prédits. La fusion de ces deux spectroscopies par les régressions MB-PLS et S-PLS conduit à des valeurs de RMSEP légèrement plus faibles : respectivement, 0,97 %(m/m) et 0,94 %(m/m). Il n'a néanmoins pas pu être démontré que cette amélioration est significative.

Pour la prédiction des teneurs en saturés, les RMSEP obtenues par les modèles MIR et PIR sont respectivement de 2,65 %(m/m) et 1,68 %(m/m). Le test de comparaison a démontré que le modèle PIR conduit à des erreurs de prédiction significativement plus faibles que le modèle MIR (*p-value* $\leq$ 0,05). Lors de la fusion des données spectrales, le modèle MB-PLS a été calculé en fixant le nombre de facteurs à 9 et une erreur de prédiction de 1,79 %(m/m) est obtenue. La RMSEP est de 1,87 %(m/m) pour le modèle S-PLS. Pour ce modèle, le premier bloc est attribué aux spectres PIR et le second aux spectres MIR. L'équation d'étalonnage est obtenue avec 8 et 5 facteurs pour exploiter les spectres PIR et MIR respectivement. Le test de comparaison "randomisation t-test" a démontré que la fusion de données par les régressions MB-PLS et S-PLS permettait d'obtenir de meilleurs résultats que le modèle MIR. Les valeurs de *p-value* montrent également que le modèle MB-PLS est équivalent au modèle PIR. Néanmoins, la qualité des prédictions obtenue

par le modèle S-PLS est moins bonne que celle du modèle PIR. Enfin, aucune différence significative n'a été observée entre les modèles MB-PLS et S-PLS.

Les résultats obtenus pour la comparaison des différentes approches dans le cadre de la prédiction de la teneur en aromatiques sont similaires à ceux de la détermination en saturés. En effet, il a été établi que la qualité des prédictions du modèle PIR est meilleure que celle du modèle MIR. Les modèles MB-PLS et S-PLS amènent également à de meilleures performances que le modèle MIR. Nous pouvons cependant noter que l'erreur de RMSEP du modèle S-PLS est plus faible que celle du modèle PIR : 1,24 %(m/m) contre 1,47 %(m/m). Le test de comparaison n'a néanmoins pas mis en évidence de différence significative entre les erreurs de prédictions de ces deux modèles (*p-value* de 0,19). Enfin, il a été démontré que la qualité des prédictions obtenues par le modèle MB-PLS est inférieure à celle des modèles PIR et S-PLS. On peut néanmoins constater que le modèle S-PLS est moins parcimonieux car 15 et 4 facteurs PLS sont nécessaires pour exploiter les bases PIR et MIR respectivement.

Deux cas de figures peuvent être distingués dans cette étude. D'une part, pour la détermination des teneurs en saturés et en aromatiques, les prédictions obtenues par les modèles MIR sont significativement plus élevées que celles des modèles PIR. La fusion de données spectrales conduit à une amélioration significative du pouvoir prédictif des modèles par rapport aux modèles MIR. De plus, les modèles MB-PLS et S-PLS peuvent conduire à des performances équivalentes aux modèles PIR. Cependant, aucune amélioration n'a pu être démontrée par l'approche de fusion par rapport aux modèles PIR. Nous pouvons également noter que les spectres PIR sont considérés en premier bloc en S-PLS pour ces deux propriétés. Comme évoqué dans la Partie 3.4.2, le premier bloc contribue généralement plus fortement à la description de la propriété et cela se traduit par une variance expliquée sur $\mathbf{y}$ et des coefficients de régression plus importants. Etant donné que la spectroscopie PIR permet d'obtenir de meilleures performances, ce constat paraît donc logique.

Le deuxième cas de figure rencontré correspond aux modèles de prédiction des teneurs en résines et en asphaltènes. Pour ces deux propriétés, les modèles MIR et PIR sont équivalents. Les RMSEP obtenues par les modèles MB-PLS et S-PLS sont, numériquement

parlant, plus faibles que celles des modèles MIR et PIR. Nous pouvons également noter que ce sont les spectres MIR qui sont considérés en premier dans les modèles S-PLS. Enfin, il a été établi que la fusion des deux bases spectrales par la MB-PLS conduisait à une amélioration significative des prédictions de la teneur en résines.

De ce constat, nous pouvons donc conclure que la spectroscopie MIR est moins adaptée pour la prédiction des teneurs en saturés et en aromatiques que la spectroscopie PIR. Afin de compléter cette étude de comparaison des spectroscopies MIR et PIR, les résultats obtenus pour la détermination de quatre autres propriétés sont indiqués en Annexe B (la densité, la teneur en hydrogène, la teneur en carbones insaturés et la teneur en carbone Conradson). Pour la détermination de la teneur en carbones insaturés, le test de comparaison a montré que les erreurs de prédiction du modèle PIR sont significativement plus faibles que celles du modèle MIR. Pour les autres propriétés considérées (densité, teneurs en hydrogène et en carbone Conradson), les équations d'étalonnages déterminées à partir de ces deux techniques sont équivalentes en termes de pourvoir prédictif. La spectroscopie PIR est donc plus performante pour la détermination des teneurs en saturés, en aromatiques et en carbones insaturés. De plus, cette technique est équivalente à la spectroscopie MIR pour les autres propriétés.

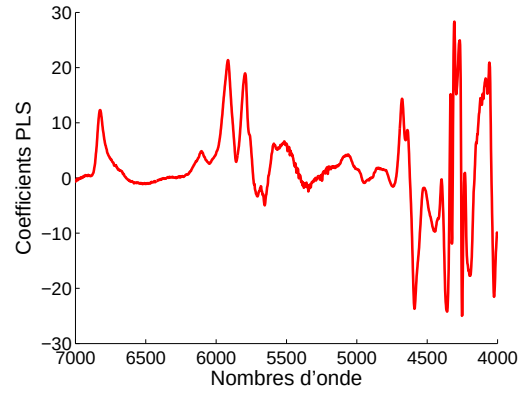
La spectroscopie MIR semble en revanche contenir de l'information qui permet de décrire les variations des teneurs en résines et en asphaltènes. Les résultats obtenus confirment que les spectroscopies MIR et PIR contiennent des informations spécifiques qui permettent une amélioration de la qualité des prédictions obtenue lors de leur exploitation simultanée.

4.3.3 Interprétation des résultats

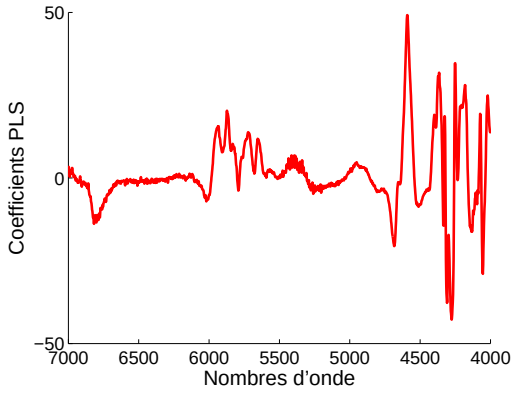
Nous commencerons dans cette partie par une interprétation des coefficients de régression des modèles MIR et PIR. Cette étude visera à définir les raisons pour lesquelles les performances des modèles MIR sont disparates en fonction des propriétés considérées. Dans un deuxième temps, nous essayerons d'identifier l'information propre à chacune de ces spectroscopies qui permet d'améliorer le niveau de prédiction pour la teneur en résines.

La Figure 4.13 représentent les coefficients de régression PLS des modèles PIR de

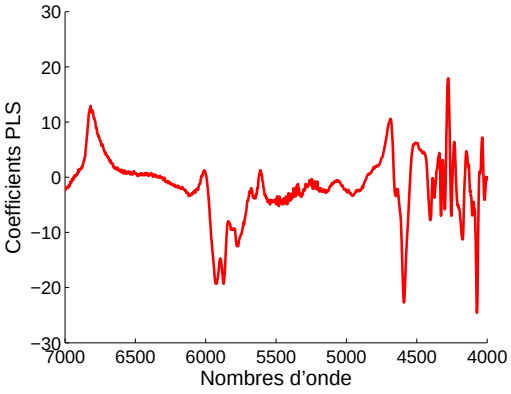
prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7. Nous pouvons tout d'abord observer que les régions spectrales où les coefficients sont les plus importants correspondent aux principales bandes qui ont été attribuées dans la Partie 4.1.2.



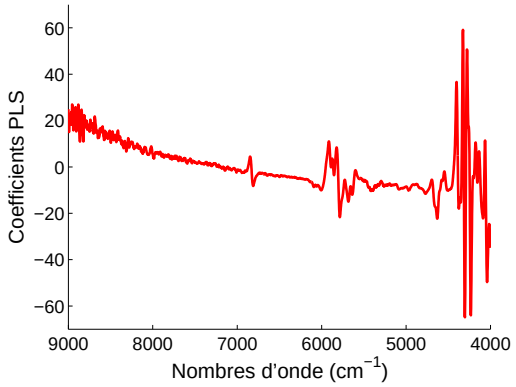
(a) Coefficients du modèle PIR de prédiction des teneurs en saturés



(b) Coefficients du modèle PIR de prédiction des teneurs en aromatiques



(c) Coefficients du modèle PIR de prédiction des teneurs en résines



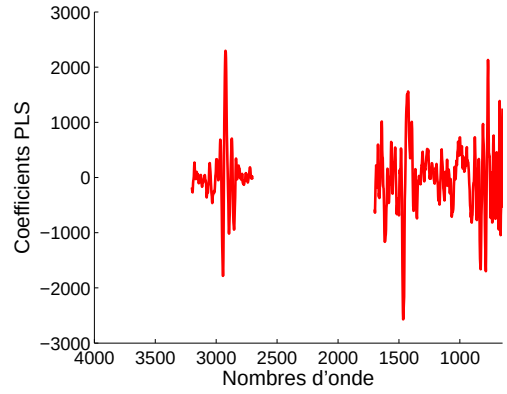
(d) Coefficients du modèle PIR de prédiction des teneurs en asphaltènes C7

FIGURE 4.13 – Coefficients de régression PLS des modèles PIR

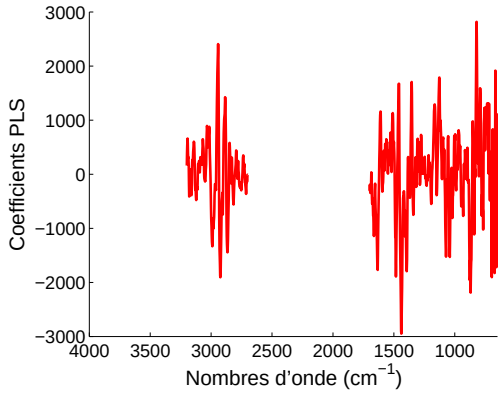
Si l'on s'intéresse aux zones spectrales attribuées aux vibrations de liaisons dans les groupements chimique saturés, on peut remarquer que les coefficients des modèles PIR de prédiction des teneurs en saturés et aromatiques sont inversés sur le domaine 4460-4000 cm^{-1} (Figures 4.13a et 4.13b). De plus, ils sont plus faibles sur cette zone spectrale pour le modèle de prédiction de la teneur en résines (Figure 4.13c). Sur le domaine 6500-5400 cm^{-1}, les coefficients du modèles PIR de détermination de la teneur en aromatiques sont plus faibles que ceux du modèle de prédiction de la teneur en saturés. Sur cette même

région, les valeurs de l'équation d'étalonnage de prédiction en résines sont négatives. La contribution des bandes attribuées aux groupements chimiques saturés paraît donc globalement en adéquation avec la composition chimique de ces trois fractions. Nous pouvons en revanche observer que les coefficients du modèle de prédiction de la teneur en asphaltènes C7 sont importants sur le domaine 4460-4000 cm^{-1} et positives entre 6500 et 5400 cm^{-1} (Figure 4.13d). L'influence de ces bandes est attribuée au fait que les molécules polycycliques présentent dans les agrégats d'asphaltènes sont principalement reliées par des chaînes paraffiniques (Partie 1.1.6). On constate également sur cette figure, l'importance de l'absorption électronique pour la description de la teneur en asphaltènes C7. En effet, les coefficients PLS augmentent fortement lorsque l'on se déplace vers les hauts nombres d'onde. Si l'on se focalise maintenant sur le domaine spectral 4700-4540 cm^{-1}, attribué à la combinaison entre l'élongation des doubles liaisons C=C et l'élongation des liaisons =C-H, nous pouvons constater la très forte influence de cette bande pour la prédiction de la teneur en aromatiques et en résines. Pour la détermination des teneurs en saturés, les valeurs de l'équation d'étalonnage sur ce domaine spectral sont inversées par rapport à celles du modèle de prédiction des teneurs en aromatiques.

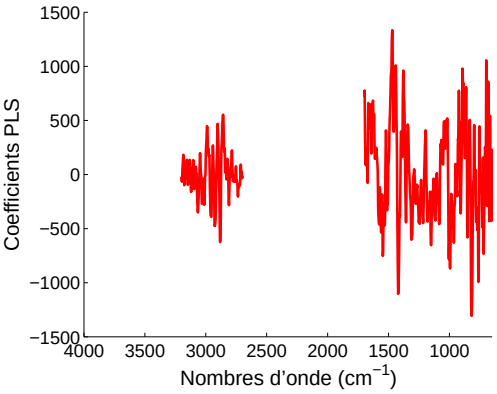
La Figure 4.14 illustre les valeurs des équations d'étalonnage des modèles MIR de prédiction des quatre fractions considérées. Nous pouvons tout d'abord observer que les coefficients de régression des modèles MIR oscillent énormément par rapport à ceux des modèles PIR. Cet effet peut tout d'abord laisser supposer un rapport signal sur bruit insatisfaisant. Nous avons néanmoins observé dans la Partie 4.1.4 que les spectres MIR dérivés ne sont pas particulièrement bruités (Figure 4.6). De plus, nous avons indiqué à titre d'exemple la 1$^{\text{ère}}$ et la 9$^{\text{ème}}$ composante du modèle MIR de prédiction des teneurs en résines (Figures 4.14e et 4.14f). On observe que la 1$^{\text{ère}}$ composante a une allure similaire aux spectres. Sur la 9$^{\text{ème}}$ composante, nous constatons l'apparition de variations qui peuvent être attribuées à du bruit. Les valeurs des poids de ces variations ne sont cependant pas très élevées. En outre, on peut observer que les valeurs des poids peuvent être très différentes selon la composante considérée. De plus, ces variations se font sur de très faibles intervalles spectraux. Nous attribuons ainsi ces fortes variations des coefficients PLS au fait qu'ils sont calculés à partir de la combinaison des différentes composantes.



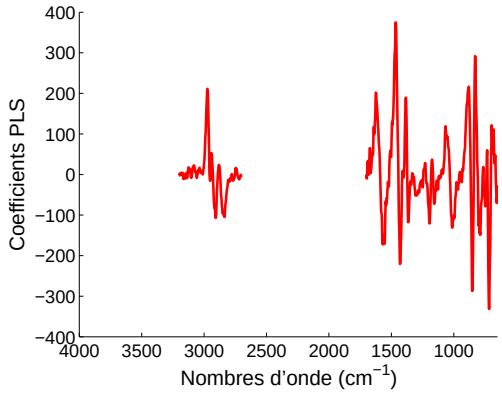
(a) Coefficients du modèle de prédiction des teneurs en saturés



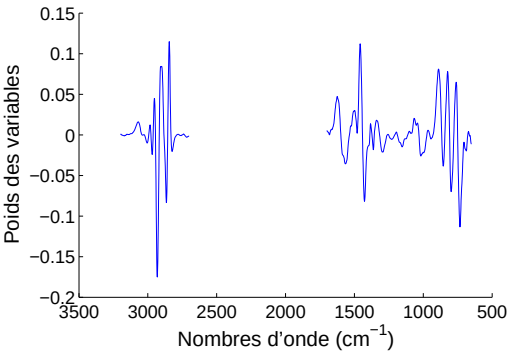
(b) Coefficients du modèle de prédiction des teneurs en aromatiques



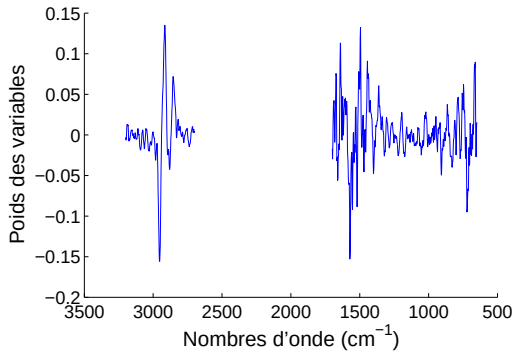
(c) Coefficients du modèle de prédiction des teneurs en résines



(d) Coefficients du modèle de prédiction des teneurs en asphaltènes C7



(e) Loadings de la 1ère composante du modèle MIR de prédiction des teneurs en résines



(f) Loadings de la 9ème composante du modèle MIR de prédiction des teneurs en résines

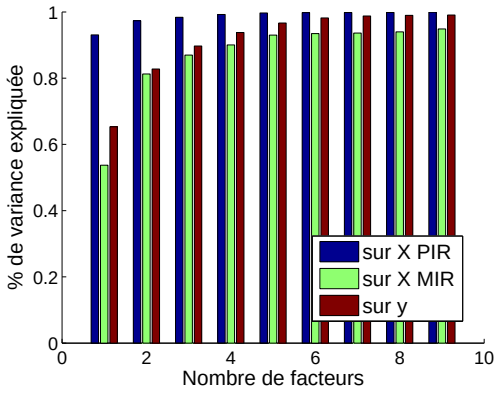
FIGURE 4.14 – Coefficients de régression PLS des modèles MIR

Une interprétation des coefficients MIR peut tout de même être effectuée. Premièrement, nous nous intéressons aux élongations des liaisons C-H dans les groupements CH_2 et CH_3 (3000-2750 cm^{-1}). On remarque que les coefficients des modèles MIR de prédiction des teneurs en saturés et en aromatiques sont importants sur cette zone spectrale. En revanche, ils sont beaucoup moins importants pour les modèles MIR de prédiction des teneurs en résines et en asphaltènes C7. Deuxièmement, on observe une forte influence des vibrations de déformation des liaisons C-H (1500-1300 cm^{-1}) pour la prédiction de ces quatre fractions. Enfin, les coefficients des bandes attribuées aux vibrations dans les groupements chimiques insaturés (950-700 cm^{-1}, 1650-1550 cm^{-1} et 3100-3000 cm^{-1}) ont une influence croissante lorsque l'on passe du modèle de prédiction de la teneur en saturés au modèle de détermination de la teneur en asphaltènes C7.

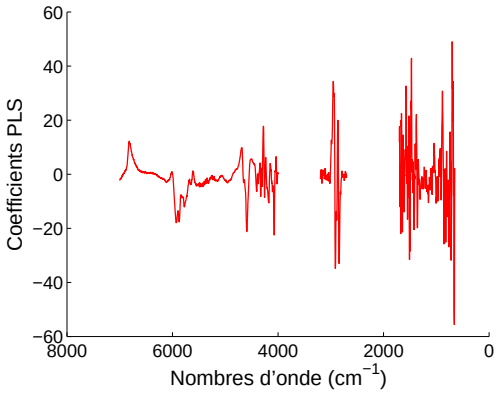
Nous constatons néanmoins que les coefficients de régression des modèles MIR de prédiction des teneurs en saturés et en aromatiques sont très élevés (-3000 à 3000) comparés à ceux des modèles PIR (-50 à 50 maximum). Il en résulte qu'une petite variation aléatoire d'intensité lors de l'acquisition du spectre peut avoir un fort impact sur les prédictions. Cette variation peut être par exemple liée à la fidélité de la mesure spectrale. Or, nous avons mentionné dans la Partie 3.2.3 que l'acquisition des spectres MIR est moins répétable que celles des spectres PIR. Cette fidélité inférieure combiné à des coefficients PLS très élevés peut être une explication à la différence entre la qualité des prédictions des modèles MIR et PIR. Les valeurs des équations d'étalonnage des modèles MIR de prédiction des résines et des asphaltènes varient également beaucoup. De plus, ces valeurs restent importantes comparées à celles des modèles PIR. On peut tout de même noter que les coefficients des modèles MIR de prédiction des résines et des asphaltènes C7 sont respectivement deux fois (-1500 à 1500) et huit fois (-400 à 400) moins forts que ceux des modèles de détermination des valeurs en saturés et en aromatiques. Nous estimons toutefois que cette diminution des valeurs des équations d'étalonnage n'est pas la seule explication à cette différence entre les performances des modèles MIR suivant les propriétés considérées. Nous allons donc maintenant nous intéresser plus particulièrement au modèle MB-PLS de prédiction des teneurs en résines. En effet, ce modèle de fusion des données spectrales a permis d'améliorer significativement le pouvoir prédictif comparati-

vement aux modèles MIR et PIR. Son interprétation devrait alors permettre de mettre en évidence les informations spécifiques de chacune des deux techniques spectroscopiques.

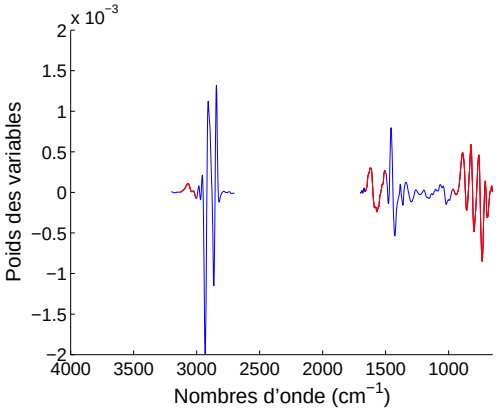
La Figure 4.15 illustre une partie des résultats obtenus pour ce modèle. Nous pouvons tout d'abord observer que les coefficients des variables PIR sont très similaires à ceux du modèle PLS développés sur les spectres PIR seuls (Figure 4.13c). En revanche, les coefficients du domaine MIR sont différents et beaucoup moins élevés que ceux du modèle MIR. En effet, ils sont du même ordre de grandeur que les coefficients PIR. La Figure 4.15a illustre le pourcentage cumulé de variance expliquée en fonction du nombre de facteurs. Cette figure montre que le pourcentage de variance expliquée par le modèle sur les spectres PIR est très important dès la première composante (93%) et qu'il augmente très faiblement par la suite pour atteintre 99,9%. Le pourcentage de variance expliquée sur les spectres MIR est relativement faible pour la première composante (53%). Il est cependant de 28% pour la deuxième composante et non négligeable pour les troisième et quatrième composantes (5 et 3% repectivement). Les constats effectués sur les variances expliquées par le modèle et sur les valeurs des coefficients de régression, nous laissent envisager que l'information présente dans le domaine MIR vient compléter l'information contenue dans la spectroscopie PIR. Une interprétation détaillée des quatre premières composantes sur le domaine MIR a donc été réalisée (Figures 4.15c à 4.15f). Sur ces quatre composantes, on observe l'importance des bandes de vibration d'élongation des liaisons carbone-hydrogène dans les groupements CH_2 et CH_3 (3000-2750 cm^{-1}). Nous pouvons néanmoins remarquer la grande influence des zones spectrales attribuées aux groupements insaturés (indiquées en rouge), notamment sur la première composante.



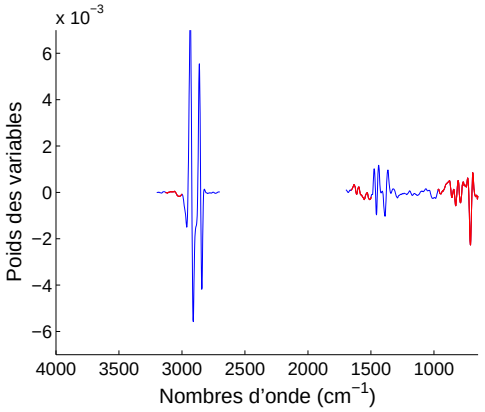
(a) Variance expliquée sur la variance totale des deux bases (X), sur les spectres PIR (X PIR) sur les spectres MIR (X MIR) et sur y (y)



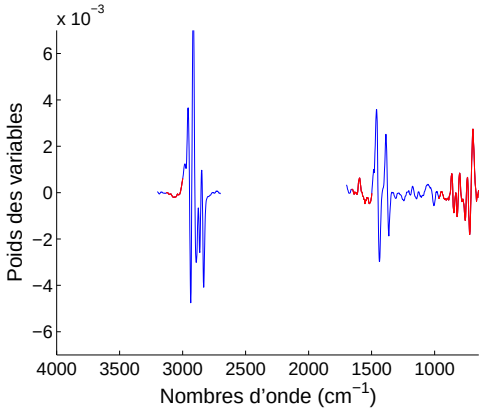
(b) Coefficients de la régression MB-PLS



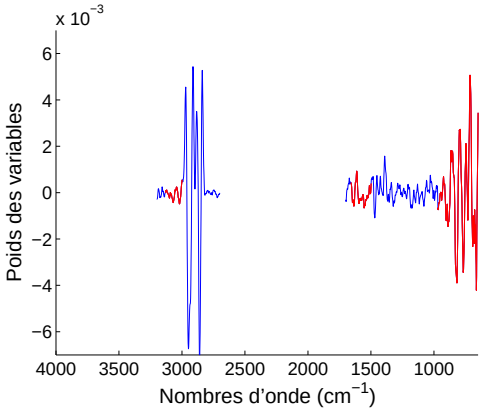
(c) Poids des variables MIR sur la 1ère composante



(d) Poids des variables MIR sur la 2ème composante



(e) Poids des variables MIR sur la 3ème composante

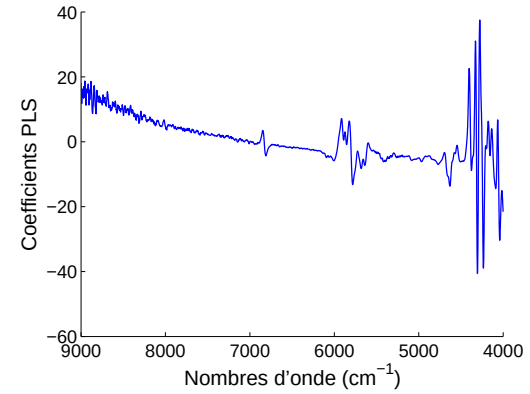


(f) Poids des variables MIR sur la 4ème composante

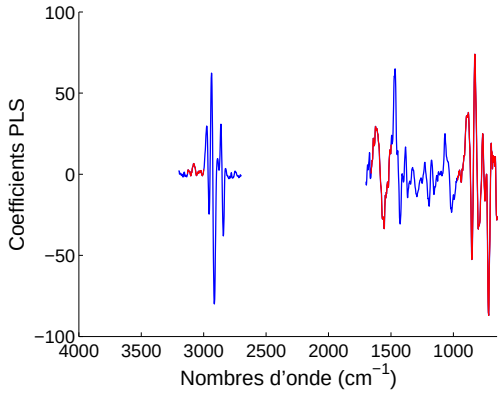
FIGURE 4.15 – Interprétation du modèle MB-PLS de prédictions des teneurs en résines

De plus, les coefficients de la bande de vibration des liaisons carbone-hydrogène hors du plan restent très importants sur les composantes 2, 3 et 4. Ces résultats montrent alors l'importance de ces trois bandes de vibration pour la description des résines et notamment l'influence de la vibration de déformation des liaisons C-H hors du plan. Nous pouvons donc supposer que cette information est déterminante dans l'amélioration des prédictions lors de la fusion des techniques spectroscopiques.

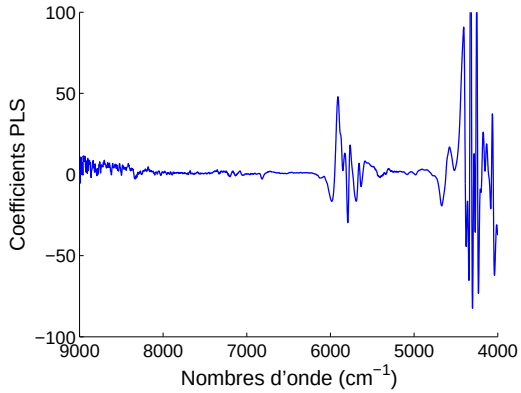
Pour finir, nous voulons également nous concentrer sur modèles MB-PLS et S-PLS de prédiction de la teneur en asphaltènes C7 qui sont également révélateurs de l'influence des bandes de vibration fondamentale des liaisons dans les groupements insaturés. Les Figures 4.16a et 4.16b illustrent l'équation d'étalonnage sur les domaines MIR et PIR du modèle MB-PLS. Comme pour la prédiction de la teneur en résines par MB-PLS, il a été constaté que les coefficients des variables PIR sont très similaires à celle du modèle développé sur les spectres PIR seuls. Comme escompté, une forte influence de l'absorption électronique est observée sur le domaine spectral 9000-6000 cm^{-1}. Nous pouvons également noter que les bandes dues aux vibrations de liaisons dans les groupements insaturés ont des valeurs de coefficients importantes. Le modèle S-PLS a été développé en considérant les spectres MIR en premier bloc. Trois facteurs ont été utilisés pour chaque bloc. Pour rappel, la qualité des prédictions de ce modèle est équivalente à celle des modèles MIR et MB-PLS. Ce modèle est très intéressant car nous pouvons constater que les coefficients PIR sur le domaine 9000-6000 cm^{-1} sont très faibles voire nuls (Figure 4.16c). L'information contenue dans l'absorption électronique, qui est une information spécifique aux asphaltènes, a donc été expliquée par les spectres MIR. Sur les coefficients des spectres MIR (Figure 4.16d), on peut effectivement constater que les bandes de vibrations des liaisons carbone-hydrogène hors du plan (900-700 cm^{-1}) et la bande d'élongation des doubles liaisons carbone-carbone ont une très forte influence. Ce résultat démontre alors toute l'importance de ces deux bandes pour la description des molécules à fort caractère aromatique.



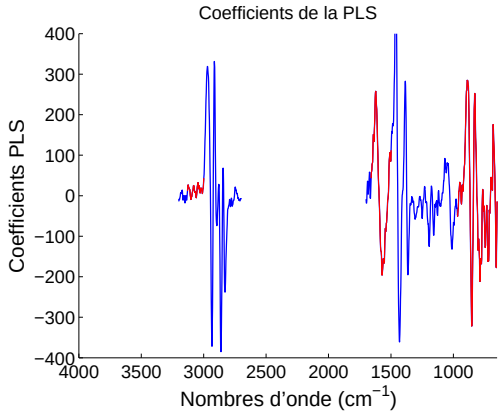
(a) Coefficients sur le domaine PIR du modèle MB-PLS



(b) Coefficients sur le domaine MIR du modèle MB-PLS



(c) Coefficients sur le domaine PIR du modèle S-PLS



(d) Coefficients sur le domaine MIR du modèle S-PLS

FIGURE 4.16 – Modèle MB-BLS et S-PLS de prédiction des teneurs en asphaltènes C7

4.3.4 Bilan

L'étude réalisée nous a tout d'abord apporté des réponses sur le potentiel des spectroscopies MIR et PIR, dans le cadre de notre application, pour la prédiction des propriétés des produits lourds. En effet, les erreurs de prédiction des modèles PIR pour la détermination des teneurs en saturés, en aromatiques et en carbones insaturés sont significativement plus faibles que celles obtenues par les modèles MIR. Aucune différence significative de la qualité des prédictions obtenues par ces deux techniques spectroscopiques n'a pu être démontrée pour les autres propriétés. Nous en avons donc déduit que la spectroscopie

PIR est globalement plus performante pour l'analyse développée au cours de ce travail de thèse.

Nous avons ensuite évalué l'apport de l'exploitation simultanée des spectroscopies MIR et PIR par les régressions MB-PLS et S-PLS. D'une part, une différence significative entre les erreurs de prédiction des modèles MIR et PIR a été démontrée pour la détermination des teneurs en saturés et en aromatiques. Les performances des modèles MB-PLS et S-PLS sont meilleures que celles des modèles MIR pour ces deux propriétés. Nous n'avons en revanche pas pu démontrer d'apport en termes de qualité de prédiction de l'exploitation simultanée par rapport aux modèles PIR. On peut donc en déduire qu'une amélioration globale n'est pas obtenue lors de la fusion des données spectrales lorsque le pouvoir prédictif des deux modèles est significativement différent. D'autre part, il a été établi que les modèles développés séparément à partir des spectroscopies MIR et PIR sont équivalents pour les teneurs en résines et en asphaltènes. Les valeurs de RMSEP obtenues par les modèles MB-PLS et S-PLS pour ces deux propriétés sont inférieures à celles obtenues par les modèles MIR et PIR. De plus, une amélioration significative a été obtenue par le modèle MB-PLS pour la détermination des teneurs en résines. Nous pouvons donc conclure que les niveaux de prédiction des modèles développés séparément à partir des techniques spectroscopiques doivent être du même ordre de grandeur pour espérer une amélioration globale du pouvoir prédictif lors de la fusion de données spectrales.

Enfin, une interprétation des modèles a été réalisée. Nous avons tout d'abord déduit de cette étude que l'information présente dans le domaine MIR vient compléter celle du domaine PIR. Nous avons ensuite recherché à identifier l'information spécifique présente dans les spectres MIR. Lors de l'interprétation des modèles MB-PLS et S-PLS, nous avons pu constater l'importance de trois bandes fondamentales pour la description des teneurs en résines et en asphaltènes : la déformation des liaisons C-H hors du plan, l'élongation des doubles liaisons C=C et, dans une moindre mesure, l'élongation des liaisons =C-H. Nous en avons conclu que ces bandes, et plus particulièrement la bande de la déformation des liaisons C-H hors du plan, semblent être à l'origine de l'amélioration des prédictions lors de la fusion de données.

4.4 Conclusions

Ce chapitre a permis de présenter les résultats qui ont été obtenus pour l'optimisation des calculs des équations d'étalonnage dans le cadre de la prédiction des teneurs en saturés, en aromatiques, en résines et en asphaltènes C7.

Nous avons tout d'abord évalué l'apport des AG pour l'optimisation simultanée du choix du prétraitement et des variables à sélectionner. Cette étude a démontré le fort potentiel cette procédure. En effet, l'optimisation du prétraitement est une solution très rapide pour sonder un grand nombre de combinaisons de méthodes et ainsi, pour définir efficacement la ou les techniques conduisant au meilleur pouvoir prédictif des modèles. Nous avons également obtenu des résultats encourageants lors de l'optimisation simultanée du prétraitements et des variables à sélectionner. Enfin, nous avons exposé le potentiel d'interprétation lié à la sélection de variables.

Nous avons par la suite procédé à une comparaison des spectroscopies MIR et PIR pour le développement de l'analyse des produits lourds. D'après les résultats obtenus, il apparaît que, pour notre application, la spectroscopie PIR est globalement plus performante que la spectroscopie MIR. En effet, une différence significative en termes de capacités de prédiction a été démontrée pour trois propriétés.

Pour finir, l'étude sur l'apport de la fusion de données a démontré le potentiel de cette approche de régression par les méthodes MB-PLS et S-PLS. Les résultats obtenus nous laissent supposer que la condition nécessaire pour espérer obtenir une amélioration du pouvoir prédictif lors de la fusion de deux spectroscopies est que les modèles développés séparément à partir de ces deux techniques amènent à des performances similaires. Enfin, les méthodes MB-PLS et S-PLS ont confirmé leur potentiel d'interprétation.

5 Caractérisation globale des produits lourds par spectroscopie PIR

Le chapitre précédent a permis de présenter les recherches menées pour l'optimisation du développement de modèles d'étalonnage multivarié. Il avait été décidé de n'effectuer ce travail que sur les déterminations des teneurs en saturés, en aromatiques, en résines et en asphaltènes C7. D'autres propriétés ont tout de même fait l'objet de développement de modèles prédictifs afin de répondre aux besoins de caractérisation globale des produits lourds du pétrole. Ces propriétés d'intérêt pour l'analyse des produits lourds sont la densité, la viscosité, la teneur en carbone Conradson et les teneurs en éléments (carbone, carbones insaturés, hydrogène, soufre et azote). Nous rappelons que les méthodes de référence pour la détermination de ces propriétés ont été décrites dans la Partie 1.3.

Les modèles de prédiction de ces propriétés n'ont pas fait l'objet d'une optimisation spécifique approfondie. En effet, les modèles présentés ici ont été obtenus assez tôt dans le travail de thèse. Ils ont été développés à partir de techniques chimiométriques classiques et en s'appuyant sur les résultats obtenus précédemment. Nous avons donc opté pour le développement de modèles à partir de la spectroscopie PIR. En effet, il a été abordé dans le chapitre précédent que cette technique spectroscopique est globalement plus performante pour l'analyse développée au cours de ce travail. Nous présenterons donc dans ce chapitre les modèles développés par spectroscopie PIR pour la prédiction de l'ensemble des propriétés d'intérêt des produits lourds. La description de l'ensemble de ces modèles peut apparaître répétitive. Elle nous semble néanmoins nécessaire pour donner, d'une

part, une vision globale de l'analyse rapide des produits lourds par spectroscopie PIR et, d'autre part, pour présenter l'ensemble du travail effectué au cours de cette thèse.

Les performances des modèles de détermination de la densité, des teneurs en carbone Conradson, en carbones insaturés et en hydrogène ont été jugées satisfaisantes pour une utilisation en laboratoire. Les corrélations obtenues pour la détermination des teneurs en soufre et en azote permettent de donner une indication mais les erreurs de prédictions semblent trop élevées pour remplacer les méthode de référence. Les résultats obtenus lors du développement des modèles prédictifs de ces six propriétés seront donc exposés dans ce chapitre.

Les performances de certains modèles sont néanmoins insuffisantes pour envisager leur utilisation au laboratoire. En effet, le nombre d'échantillons disponibles est trop faible pour le développement d'un étalonnage multivarié de la teneur en carbone et la teneur en asphaltènes. De plus, une non-linéarité entre les valeurs de la viscosité et les spectres PIR a été détectée. Plusieurs approches ont été testées pour pallier ce problème mais sans résultat. Ces modèles sont tout de même décrits en Annexe B.2 pour illustrer les difficultés du développement de modèles avec ce genre de problèmes.

5.1 Description général des modèles développés

Les modèles de prédiction qui seront présentés par la suite ont été développés en utilisant la régression PLS. Le nombre de facteurs PLS a été fixé en s'appuyant sur les erreurs de RMSECV obtenues en validation croisée. Le lot de validation représente 25% des échantillons disponibles et a été sélectionné en fonction des valeurs des propriétés. Enfin, la méthode *bootstrap* a été appliquée afin d'encadrer les valeurs prédites par un niveau de confiance.

Nous présenterons, pour chaque modèle, les résultats obtenus en validation croisée pour la détermination du nombre de facteurs optimal, le pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ ainsi que les coefficients de régression PLS. Les graphiques des résidus de prédiction des échantillons du lot de validation et le graphique de parité seront également exposés. Les échantillons des lots d'étalonnage y sont indiqués afin d'illustrer la

gamme analytique considérée. Les niveaux de confiance calculés par la méthode *bootstrap* seront également tracés pour l'encadrement de chaque valeur prédite. Il faut noter que l'intervalle de confiance (IC) de la méthode de référence sera également indiqué sur ces figures afin de donner un élément de comparaison de la qualité des prédictions. Le Tableau 5.1 synthétise les gammes analytiques considérées ainsi que le domaine spectral et le prétraitement appliqués pour le développement des étalonnages multivariés. Le pourcentage d'échantillons aberrants qui ont été retirés de la base est également indiqué. Enfin, le nombre de facteurs utilisés et la RMSEP obtenue sont présentés.

Le Tableau 5.1 illustre que la plupart des modèles ont été développés sur le domaine 7000-4000 cm^{-1} en appliquant la méthode WLSB avec un polynôme d'ordre 3. En effet, ces paramètres permettent de s'affranchir des variations liées à l'absorption électronique. En revanche, le modèle de prédiction de la teneur en carbone conradson a été développé sur les spectres dérivés sur le domaine 9000-4000 cm^{-1}. Ces paramètres seront justifiés lors de la description de ce modèle.

Tableau 5.1 – Résultats du développement de l'analyse multivariée des produits lourds par spectroscopie PIR

Propriété	Gamme analytique	Domaine spectral (cm^{-1})	Prétraitement	% d'échantillons aberrants	Nombre de facteurs PLS	RMSEP
Densité	0,8186 - 1,0243	7000 - 4000	WLSB (3)[a]	6%	6	0,0035
Viscosité	3,45 - 212,60 cSt	9000 - 4000	Dérivée 1ère	1%	7	24 cSt
Saturés	14,1 - 99,1 %(m/m)	7000 - 4000	WLSB (3)[a]	7%	8	1,68 %(m/m)
Aromatiques	0,7 - 61,1 %(m/m)	7000 - 4000	WLSB (3)[a]	5%	9	1,47 %(m/m)
Résines	0,1 - 34,5 %(m/m)	7000 - 4000	WLSB (3)[a]	2%	9	1,10 %(m/m)
Asphaltènes	3,5 - 12,8 %(m/m)	9000 - 4000	Dérivée 1ère	12,5%	3	0,85 %(m/m)
Asphaltènes C7	0,5 - 14,2 %(m/m)	9000 - 4000	Dérivée 1ère	2%	3	1,05 %(m/m)
Carbone Conradson	1,00 - 20,03 %(m/m)	9000 - 4000	Dérivée 1ère	3%	7	1,11 %(m/m)
Carbone	86,01 - 88,40 %(m/m)	7000 - 4000	WLSB (3)[a]	3%	5	0,5 %(m/m)
Carbones Insaturés	2,0 - 36,8 %(m/m)	7000 - 4000	WLSB (3)[a]	4%	5	1,18 %(m/m)
Hydrogène	9,95 - 14,59 %(m/m)	7000 - 4000	WLSB (3)[a]	2%	6	0,17 %(m/m)
Azote	500 - 11 800 ppm	7000 - 4000	WLSB (3)[a]	1%	10	628 ppm
Soufre	1,00 - 3,54 %(m/m)	7000 - 4000	WLSB (3)[a]	3%	6	0,28 %(m/m)

[a] : méthode *Weighted Least Square baseline* avec un polynôme d'ordre 3

5.2 La détermination des propriétés physico-chimiques globales

Les propriétés physico-chimiques globales qui ont fait l'objet du développement d'un étalonnage multivarié au cours de ce travail sont la densité et la viscosité. Nous avons cependant évoqué que le modèle de prédiction de la viscosité est insatisfaisant du fait de la présence d'une non-linéarité avec les spectres de vibration. Ce modèle est tout de même illustré dans l'Annexe B.2.1. Nous avons décidé de ne pas déterminer l'indice de réfraction car il ne s'applique pas aux produits les plus opaques. La méthode de référence de l'indice de réfraction est de plus très fidèle (reproductibilité de 6.10^{-5}). Cette analyse sert ainsi le plus souvent à différencier des produits très proches. Enfin, le temps d'analyse étant très court (25 minutes), l'analyse spectroscopique de cette propriété n'est pas prioritaire.

5.2.1 La densité

La base de données contient 222 échantillons renseignés en densité. La gamme analytique couverte par ces échantillons s'étend de 0,8186 à 1,0243. Lors du développement du modèle, 13 échantillons ont été déclarés aberrants (6% des échantillons). En effet, leurs résidus et leurs influences sur le modèle étaient trop importants.

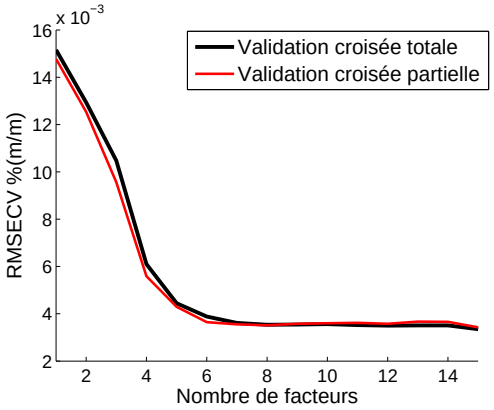
La Figure 5.1 résume les résultats obtenus pour la détermination des valeurs de densité. Tout d'abord, la Figure 5.1a illustre la courbe des valeurs de RMSECV obtenues en validation croisée totale et partielle en fonction du nombre de facteurs PLS. On peut observer dans les deux cas que l'erreur de RMSECV décroît jusqu'au sixième facteur puis admet un palier. Le modèle de prédiction de la teneur en densité a donc été développé en utilisant 6 facteurs PLS.

Les pourcentages cumulés de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ sont indiqués sur la Figure 5.1b. Nous pouvons constater que la 1ère composante explique la majorité de la variance sur $\mathbf{X}$ et sur $\mathbf{y}$ (93,3% et 91%, respectivement). Les contributions à la variance expliquée augmentent très peu ensuite pour atteindre 99,8% sur $\mathbf{X}$ et 99,5% sur $\mathbf{y}$.

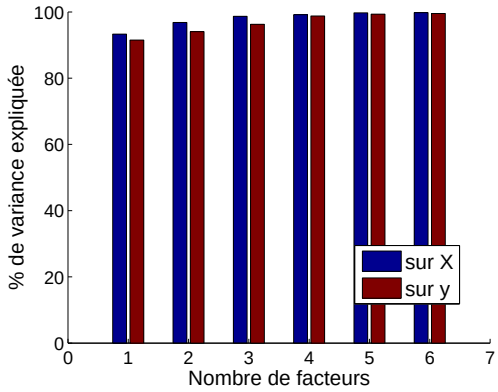
Sur la Figure 5.1c nous pouvons observer que les coefficients des zones spectrales dues aux vibrations des liaisons dans les groupements chimiques saturés sont négativement

corrélées à la teneur en densité (4460-4000 cm^{-1} et 6500-5400 cm^{-1}). En revanche, on peut remarquer que les coefficients de la bande causée par les vibrations des liaisons dans les groupements insaturés sont importants. Ces observations apparaissent logiques car le caractère aromatique des produits pétroliers augmente avec la valeur de densité.

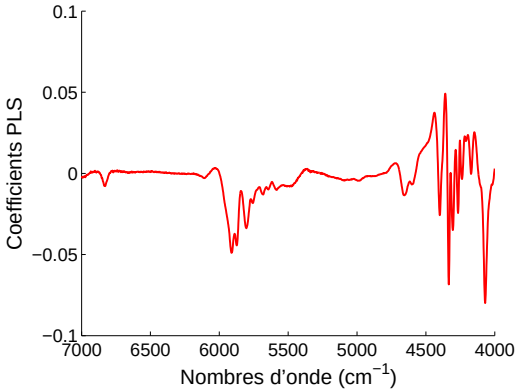
La Figure 5.1d représente les résidus en fonction de la valeur de référence. Les résidus des échantillons du lot d'étalonnage sont indiqués pour illustrer la gamme analytique couverte (en bleu). L'intervalle de confiance (IC) de la méthode de référence est également tracé afin de donner une indication sur les performances obtenues par le modèle (en vert). Enfin, les valeurs prédites par le modèle pour les échantillons du lot de validation sont représentées et encadrées par les valeurs minimales et maximales obtenues par la méthode de *bootstrap* (en rouge). Sur ce graphique, nous pouvons constater que la plupart des échantillons de validation sont prédits avec des résidus compris entre -0,005 et 0,005. De plus, nous pouvons remarquer sur la Figure 5.1e que le résidus et les encadrements des valeurs prédites sont relativement faibles comparés à la gamme analytique considérée. Une valeur de RMSEP de 0,0035 a par ailleurs été obtenue.



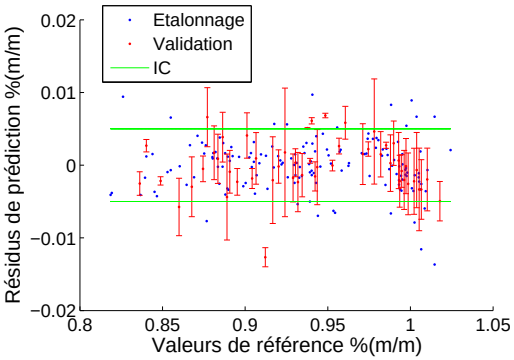
(a) Validation croisée pour le choix du nombre de facteurs PLS



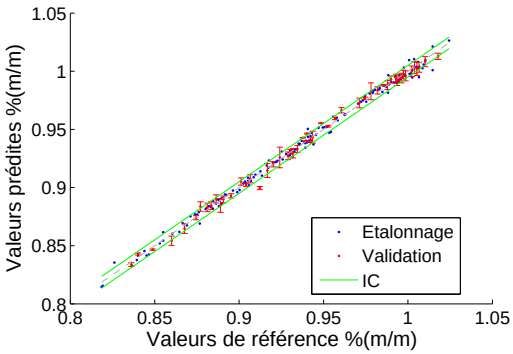
(b) Pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction



(e) Diagramme de parité

FIGURE 5.1 – Résultats pour le modèle de détermination des valeurs de la densité

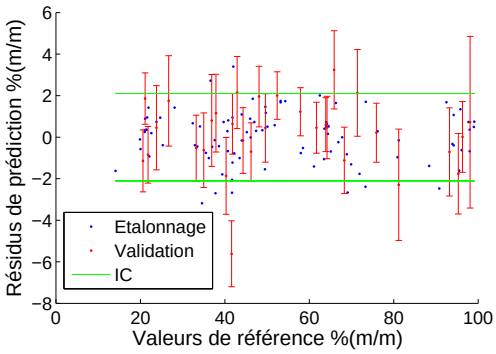
5.3 La détermination des propriétés de répartition massique par familles chimiques

Les propriétés de répartition massique par familles chimiques qui ont fait l'objet du développement d'un étalonnage multivarié au cours de ce travail sont les teneurs en SARA, en asphaltènes C7 et en carbone Conradson. Nous ne détaillerons pas les modèles de prédiction des teneurs en saturés, en aromatiques, en résines (SAR) et en asphaltènes C7, car ils ont déjà été présentés dans la Partie 4.3. Pour ces modèles, nous n'exposerons que les graphiques des résidus et les graphiques de parité pour illustrer les résultats obtenus lors de l'application de la méthode *bootstrap*. Les performances obtenus pour la détermination de la teneur en asphaltènes, obtenus par le fractionnement SARA, sont insatisfaisantes du fait du faible nombre d'échantillons disponibles. Il a donc été choisi de présenter ce modèle en Annexe B.2.2.

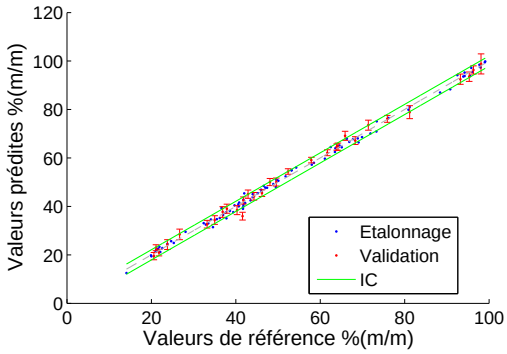
5.3.1 Les performances des modèles SAR et asphaltènes C7

Les résultats obtenus par la méthode *bootstrap* pour la prédiction des teneurs en saturés, en aromatiques, en résines et en asphaltènes C7 sont exposés ici.

L'équation d'étalonnage pour la détermination de la teneur en saturés a été développée à partir de 124 échantillons sur une gamme analytique qui couvrait de 14,1 à 99,1 %(m/m). La valeur de RMSEP obtenue est de 1,68 %(m/m). La Figure 5.2a illustre que les résidus des valeurs prédites pour les échantillons du lot de validation sont globalement compris entre -2 et 2 %(m/m). De plus, les niveaux de confiance calculés par la méthode *bootstrap* sont relativement constants sur la gamme analytique. La Figure 5.2b montre que les erreurs de prédiction et les niveaux de confiance sont faibles pour la gamme analytique considérée.



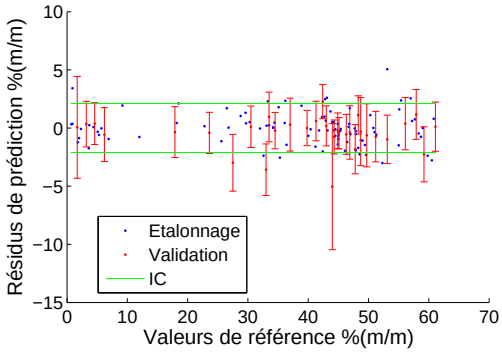
(a) Résidus de prédiction



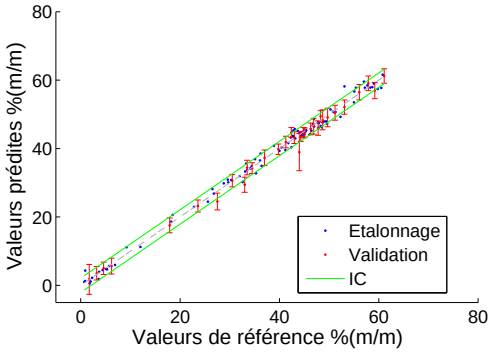
(b) Diagramme de parité

FIGURE 5.2 – Performances du modèle de détermination de la teneur en saturés

Pour la prédiction de la teneur en aromatiques, 127 échantillons sont disponibles et la gamme analytique s'étend de 0,7 à 61,1%. Les performances de ce modèle sont également satisfaisantes. En effet, la valeur de RMSEP est de 1,47 %(m/m) et les résidus sont globalement compris entre -2 et 2 (Figure 5.3a).



(a) Résidus de prédiction



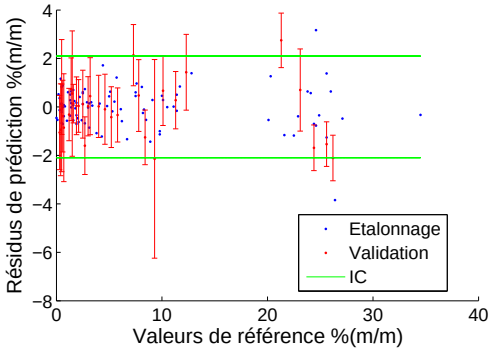
(b) Diagramme de parité

FIGURE 5.3 – Résultats pour le modèle de détermination de la teneur en aromatiques

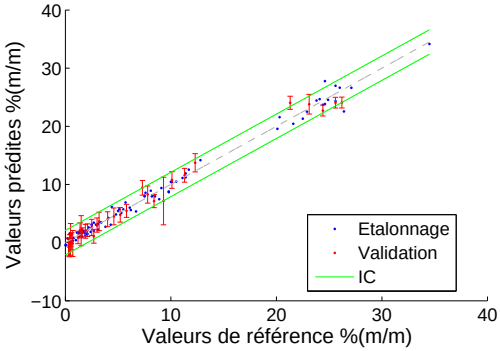
Les niveaux de confiance calculés sur les valeurs prédites ont également des amplitudes relativement faibles. Nous pouvons cependant constater que certaines valeurs sont très proches de zéro. On peut d'ailleurs remarquer que la valeur minimum du niveau de confiance est négative pour l'échantillon de validation dont la teneur en aromatiques est la plus faible (Figure 5.3b). La méthode de *bootstrap* peut ainsi servir également à identifier les prédictions critiques. Dans le cadre de l'utilisation en laboratoire, une valeur prédite

potentiellement égale à zéro serait indiquée pour cet échantillon.

Le modèle de prédiction de la teneur en résines a été développé à partir de 130 échantillons sur une gamme analytique allant de 0,1 à 34,5 %(m/m). Une valeur de RMSEP de 1,10 %(m/m) a été obtenue. La Figure 5.4a montre que l'erreur sur les prédictions des échantillons de validation sont faibles.



(a) Résidus de prédiction



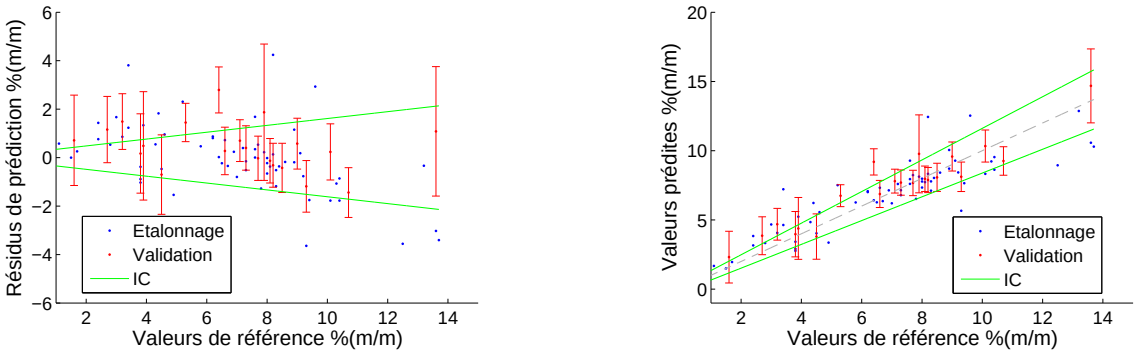
(b) Diagramme de parité

FIGURE 5.4 – Résultats pour le modèle de détermination de la teneur en résines

De plus, les amplitudes des niveaux de confiance sont également faibles. Ici encore, certaines valeurs minimales des niveaux de confiance sont négatives pour les valeurs prédites proches de zéro. Ces prédictions seront alors déclarées critiques et des valeurs prédites potentiellement égales à zéro seront indiquées. Nous pouvons également constater sur la Figure 5.4b un échantillon dont la valeur en résines est très élevée par rapport aux autres échantillons de la base. Cet échantillon, qui ne dégrade pas le modèle, est donc introduit pour le calcul de l'équation d'étalonnage. Ce modèle peut être appliqué en laboratoire sur une gamme analytique allant de 0% à 30%. Il sera cependant nécessaire d'enrichir la gamme analytique entre 30 et 35 %(m/m) pour une utilisation du modèle sur une gamme plus large.

Le modèle pour la détermination des teneurs en asphaltènes C7 a été développé à partir de 89 échantillons qui couvrent la gamme analytique 0,5 - 13,7 %(m/m). En effet, comme nous l'avons évoqué dans la Partie 4.1.1, les échantillons dont la teneur est égale à zéro n'ont pas été introduits dans le modèle. Les résultats sont satisfaisants (RMSEP = 1,05 %(m/m)). La Figure 5.5a illustre que les résidus sont compris entre -2 et 2 %(m/m).

Les amplitudes des niveaux de confiance sont néanmoins importantes mais ce modèle est relativement satisfaisant au regard de la fidélité de la méthode de référence (répétabilité $= 0,2 + 0,2 \times y$, où y représente la teneur en asphaltènes C7).



(a) Résidus de prédiction

(b) Diagramme de parité

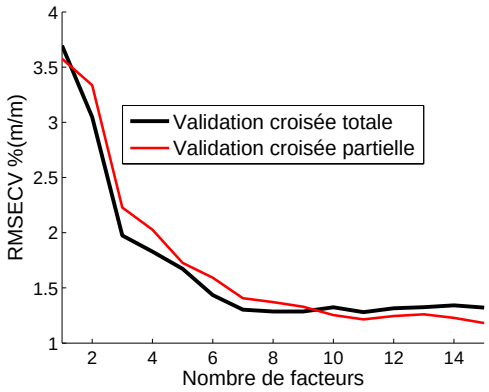
FIGURE 5.5 – Performances du modèle de détermination de la teneur en asphaltènes C7

5.3.2 La teneur en carbone Conradson

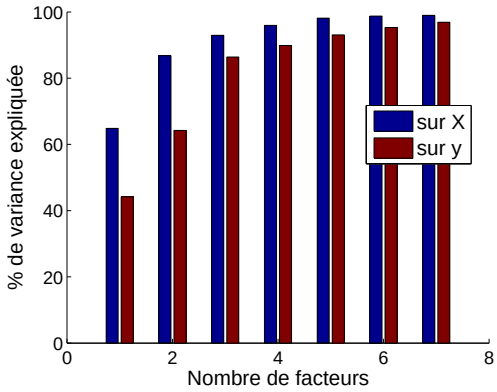
La gamme analytique de la teneur en carbone Conradson couverte par la base d'échantillons s'étend de 0,05 à 20,03 %(m/m). Il a néanmoins été remarqué que les échantillons dont la teneur en carbone Conradson est inférieure à 1 %(m/m) dégradent fortement le modèle. En effet, comme nous l'avons précédemment évoqué, les modèles de prédiction ne sont généralement pas adaptés pour la détermination des faibles valeurs. Les résidus obtenus sont alors très importants. Ces échantillons ont donc été retirés de la base. Le modèle a été développé à partir de 98 échantillons sur une gamme analytique de 1,00 à 20,03 %(m/m). Les spectres ont été prétraités en dérivée 1$^{\text{ère}}$ sur le domaine 9000-4000 cm^{-1} car l'absorption électronique présente sur le domaine 9000-6500 cm^{-1} est une information pertinente pour ce modèle. L'analyse du carbone Conradson consiste en effet à peser le résidu de produit après chauffage de l'échantillon à forte température (500°C) pendant 15 minutes. Les molécules qui composent le résidu sont les molécules dont les points d'ébullition sont les plus hauts. La plupart de ces molécules appartiennent donc à la fraction asphaltène.

La Figure 5.6a montre que la courbe des valeurs de RMSECV, obtenues par la validation croisée totale, en fonction du nombre de facteurs PLS admet un palier à partir de 7 composantes. En revanche, la courbe correspondant aux valeurs de RMSECV obtenues par validation croisée totale diminue jusqu'à la 11$^{\text{ème}}$ composante. Le nombre de facteurs PLS a été fixé à 7 car les coefficients sont déjà très bruités pour un modèle à 7 facteurs PLS (Figure 5.6c). Comme nous l'avons évoqué dans la Partie 4.1.3, des méthodes de lissage ont été appliquées. Dans le cadre de la prédiction du carbone Conradson, ces techniques permettent de réduire fortement le bruit présent sur les coefficients PLS.

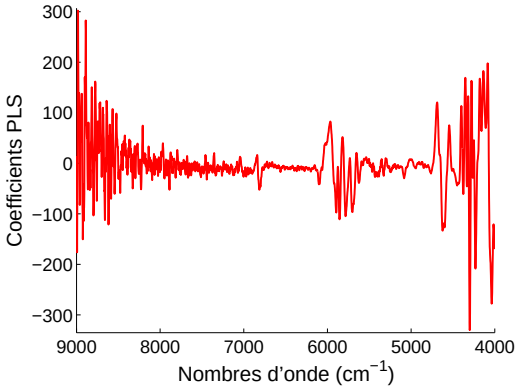
La valeur de RMSEP obtenue pour ce modèle est de 1,11 %(m/m). On peut remarquer que les résidus de prédiction se situent entre -1,5 et 1,5 %(m/m), mis à part deux échantillons dont les teneurs sont très élevées. De plus, l'amplitude maximale des niveaux de confiance calculés par la méthode *boostrap* sur les valeurs prédites est de 1,7 %(m/m).



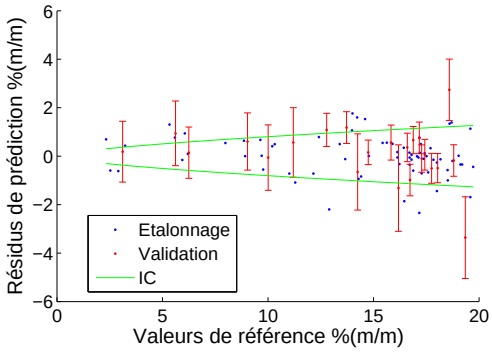
(a) Validation croisée pour le choix du nombre de facteurs PLS



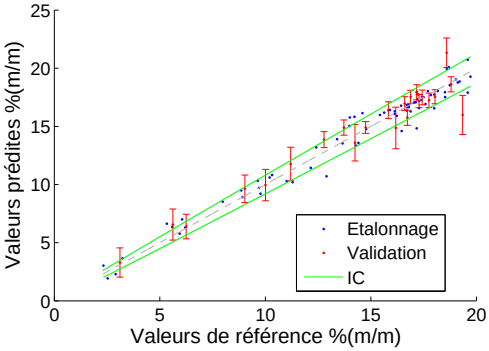
(b) Pourcentage cumulé de variance expliquée sur **X** et sur **y** en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction



(e) Diagramme de parité

FIGURE 5.6 – Résultats pour le modèle de détermination de la teneur en carbone Conradson

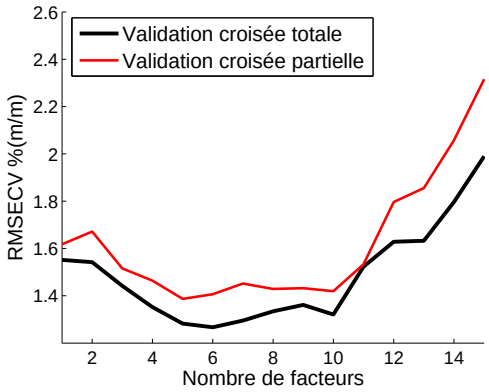
5.4 La prédiction des teneurs en éléments

Au cours de ce travail, nous avons développé des étalonnages multivariés pour la prédiction de la teneur en carbones insaturés, en hydrogène, en azote et en soufre. Afin de déterminer la faisabilité de la prédiction de la teneur en carbone, un modèle a été calculé malgré le faible nombre d'échantillons disponibles. Les résultats obtenus étant insatisfaisants, ce modèle est présenté en Annexe B.2.3.

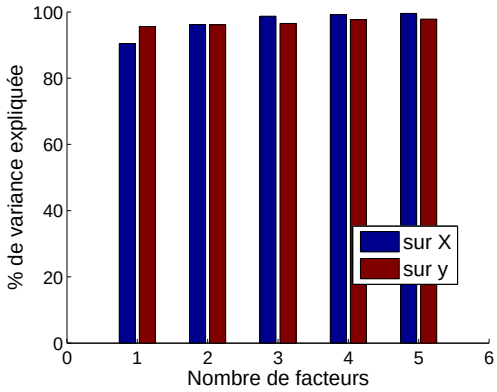
5.4.1 La teneur en carbones insaturés

Le modèle de détermination de la teneur en carbones insaturés a été développé à l'aide de 81 échantillons sur une gamme analytique qui s'étend de 2,0 à 36,8 %(m/m). L'équation d'étalonnage a été calculée à partir de 5 facteurs PLS. En effet, la courbe des valeurs de RMSECV, obtenues par validation partielle, en fonction du nombre de composantes admet un minimum pour 5 composantes (Figure 5.7a). La Figure 5.7c illustre que les coefficients PLS sont négativement corrélés sur les zones spectrales attribuées aux bandes de vibration de liaisons dans les groupements chimiques saturés (4460-4000 cm^{-1} et 6200-5400 cm^{-1}). On peut également remarquer la forte influence de la bande de combinaison des élongations des liaisons C=C et =C-H (4700-4540 cm^{-1}).

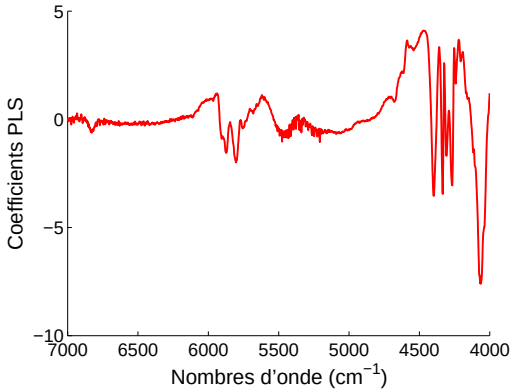
Les Figures 5.7d et 5.7e montrent que les prédictions obtenues par ce modèle sont de très bonnes qualités. Par ailleurs, une valeur de RMSEP de 1,18 %(m/m) a été obtenue. Nous pouvons tout de même noter que seulement trois échantillons du lot d'étalonnage sont présents sur la gamme 3 - 12 %(m/m). Ces échantillons ne dégradent pas les performances du modèle mais la qualité des prédictions n'a pas été évaluée sur cette gamme faute de valeurs disponibles. Nous estimons donc que ce modèle peut-être validé mais qu'il sera nécessaire d'enrichir cette gamme analytique par de nouveaux échantillons.



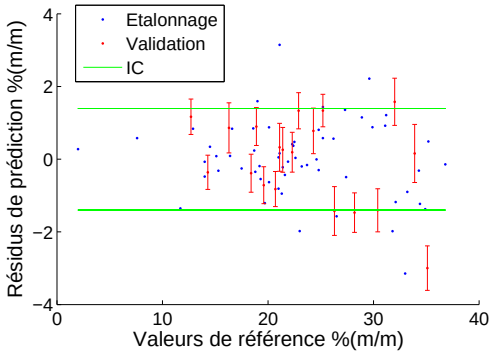
(a) Validation croisée pour le choix du nombre de facteurs PLS



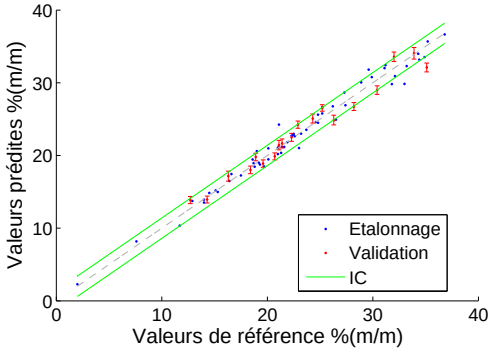
(b) Pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction

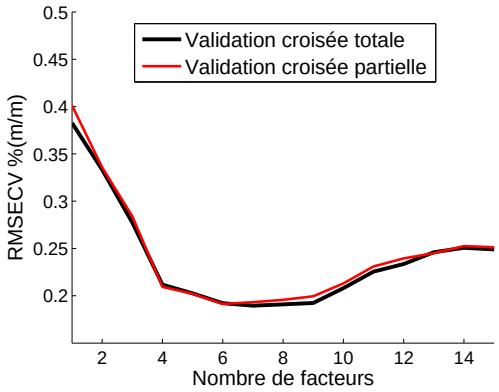


(e) Diagramme de parité

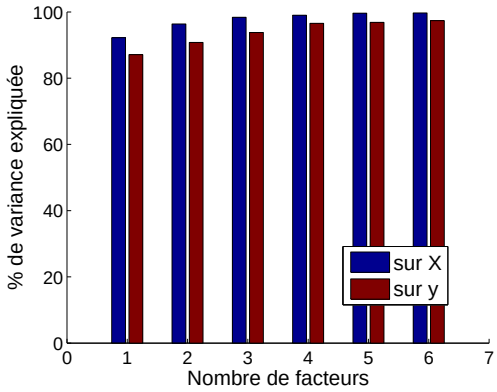
FIGURE 5.7 – Résultats pour le modèle de détermination de la teneur en carbones insaturés

5.4.2 La teneur en hydrogène

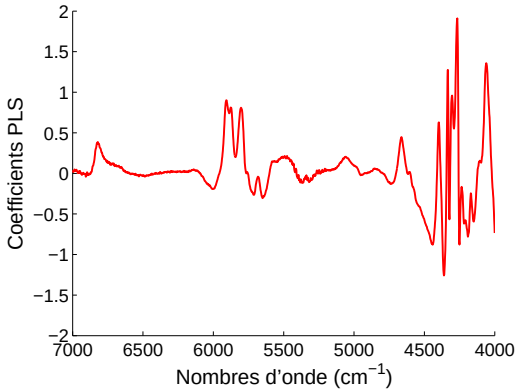
Les 171 échantillons disponibles pour la teneur en hydrogène couvre la gamme analytique 9,95 - 14,59 %(m/m). Le nombre de facteurs optimal a été fixé à 6 (Figure 5.8a). La Figure 5.8c illustre que la corrélation positive des coefficients avec les bandes situées sur les zones spectrales 4460-4000 cm^{-1} et 6200-5400 cm^{-1}. Sur le domaine 4700-4500 cm^{-1}, attribué aux vibrations de liaison dans les groupements insaturés, les coefficients sont négatifs. Ces constats paraissent en adéquation avec la propriété à déterminer. La prédiction des valeurs en hydrogène des échantillons de validation correspond à une erreur de RMSEP de 0,17 %(m/m). Les Figures 5.8d et 5.8e illustrent que les résidus de prédiction sont relativement faibles. En effet, les erreurs de prédiction sont principalement comprises entre -0,23 et 0,23 %(m/m). Nous pouvons donc conclure que cet étalonnage multivarié permet une détermination satisfaisante de la teneur en hydrogène. Nous pouvons néanmoins constater que peu de teneurs en hydrogène sont disponibles au dessus de 14 %(m/m). Il en résulte que l'amplitude du niveau de confiance de l'échantillon de validation prédit sur cette gamme est très élevée. Lors d'une future implémentation, il sera alors nécessaire d'introduire dans la base de données des échantillons dont les valeurs en hydrogène seront supérieures à 14%(m/m).



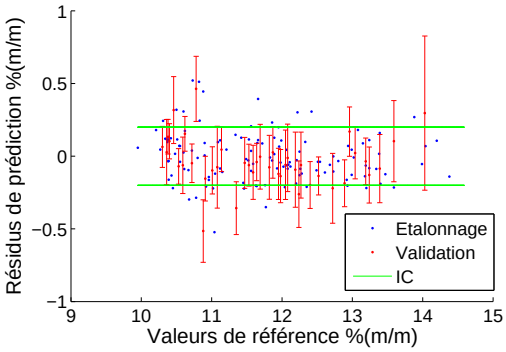
(a) Validation croisée pour le choix du nombre de facteurs PLS



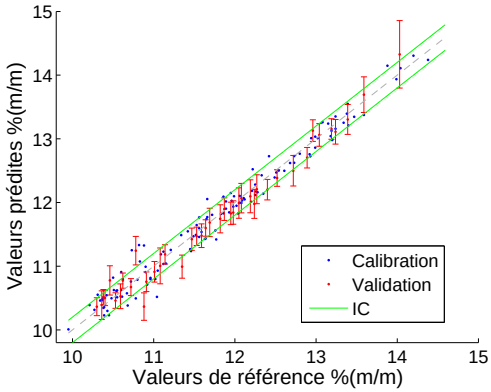
(b) Pourcentage cumulé de variance expliquée sur **X** et sur **y** en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction



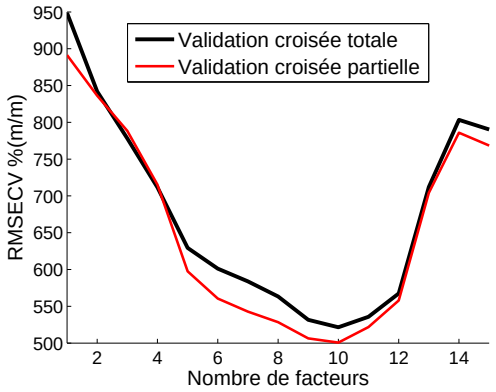
(e) Diagramme de parité

FIGURE 5.8 – Interprétation du modèle de prédiction de la teneur en hydrogène

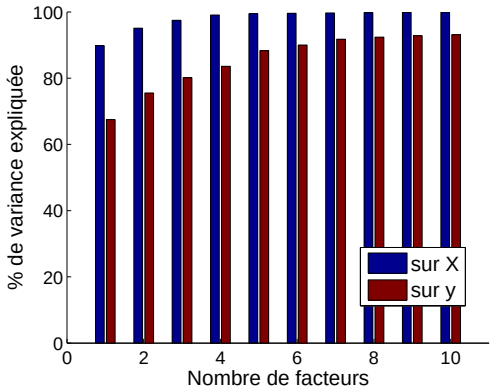
5.4.3 La teneur en azote

La gamme analytique de la teneur en azote couverte par la base d'échantillons s'étend de 0,1 à 11800 ppm. Cependant, une dégradation des modèles a été constatée lors de la prise en compte des valeurs de 0 à 500 ppm. L'étalonnage a donc été développé à partir de 108 échantillons sur une gamme analytique qui allait de 500 à 11800 ppm. Le modèle PLS a été calculé avec 10 facteurs. Nous pouvons constater sur les Figures 5.9d et 5.9e que les erreurs de prédiction sont relativement importantes. En effet, la RMSEP obtenue est de 628 ppm. De plus, les amplitudes des niveaux de prédiction sont assez élevées notamment pour deux échantillons. Cependant, cet étalonnage est plutôt satisfaisant car les teneurs considérées sont très faibles.

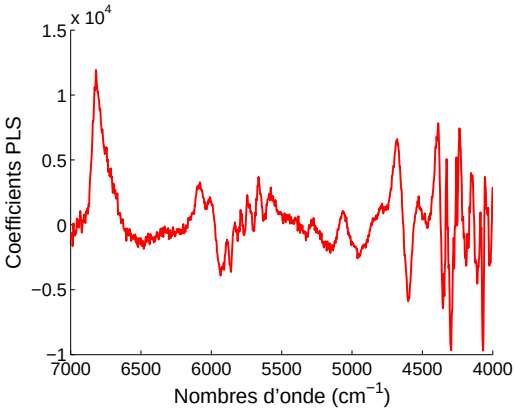
L'interprétation de ce modèle permet de constater d'une part que la variance expliquée sur $\mathbf{y}$ reste assez faible en comparaison aux autres modèles, même lorsque 10 composantes sont considérées (Figure 5.9b). D'autre part, la Figure 5.9c illustre que la bande entre 7000 et 6500 cm^{-1} a une très forte influence sur le modèle. Nous avons mentionné, lors de l'interprétation des modèles développés à partir des spectres PIR (Partie 4.3.3), qu'une variation très faible de l'absorbance peut être observée sur ce domaine spectral. Dans les travaux présentés dans le chapitre 2, aucune attribution de cette bande n'a été relevée. En outre, les tables de correspondance entre la structure chimique et les bandes d'absorption dans le domaine PIR indiquent que la première harmonique des vibrations des liaisons N-H peut être observée sur le domaine 7100-6700 (1400-1500 nm). Ces bandes sont dues à la présence de groupements RNH_2, CONHR et $CONH_2$. On peut donc supposer que les fortes valeurs des coefficients aux alentours de 6800 cm^{-1} sont dues à la première harmonique de vibration des liaisons N-H dans les groupements RNH_2. Il est néanmoins difficile de confirmer cette hypothèse.



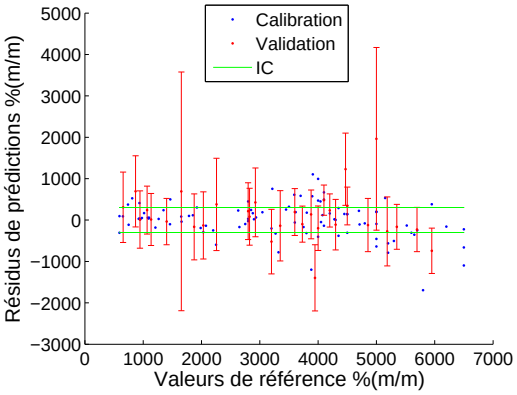
(a) Validation croisée pour le choix du nombre de facteurs PLS



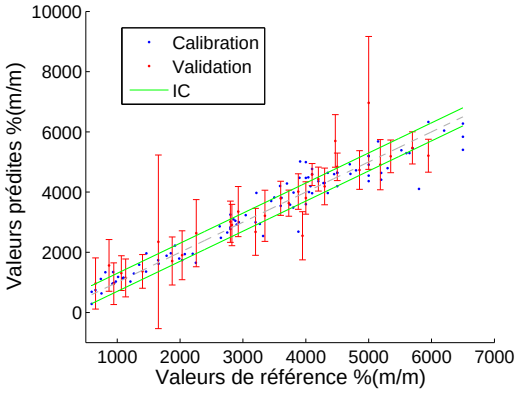
(b) Pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction



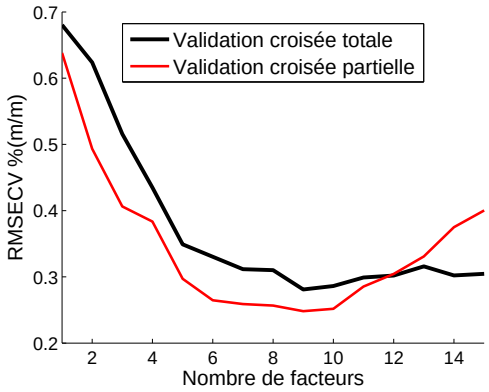
(e) Diagramme de parité

FIGURE 5.9 – Interprétation du modèle de prédiction de la teneur en azote

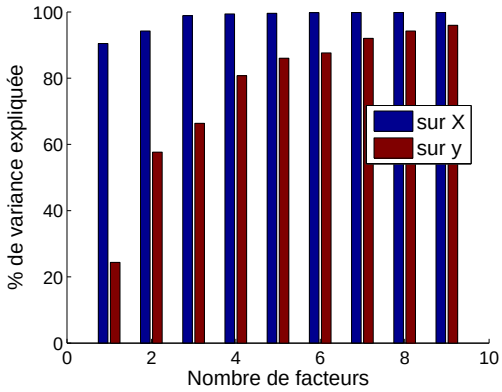
5.4.4 La teneur en soufre

De façon similaire au modèle de prédiction en azote, les échantillons à faibles teneurs en soufre (< 1%(m/m)) ont été supprimés de la base de données. Ainsi, le modèle a été développé à partir de 83 échantillons sur une gamme analytique de 1 à 3,54 %(m/m). Neuf facteurs PLS ont été utilisés pour le calcul de cet étalonnage multivarié.

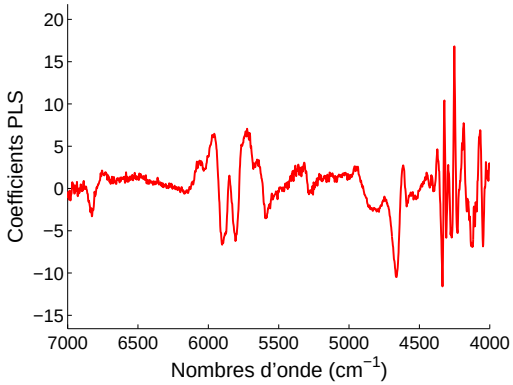
Nous pouvons constater sur la figure 5.10b que le pourcentage de variance expliquée sur **y** est plus faible pour le premier facteur (24 %) que pour le deuxième (33%). Inversement, la contribution du premier facteur à la variance expliquée sur **X** est très forte (90,5 %) et beaucoup plus faible pour le deuxième (3,7 %). Par définition, la première composante PLS explique une plus grande covariance entre **X** et **y** que la deuxième composante. Nous pouvons donc déduire que les spectres PIR et la teneur en soufre ne sont pas très corrélés. Ce modèle est néanmoins relativement satisfaisant car l'erreur de RMSEP obtenue (0,28 %(m/m)) est assez faible avec des résidus répartis entre -0.5 et 0.3 %(m/m) (Figures 5.10d et 5.10e). De plus, le pourcentage cumulé de variance expliquée sur **y** atteint tout de même 96 % à la neuvième composante. Il faut néanmoins noter que les amplitudes des niveaux de confiance sont élevées.



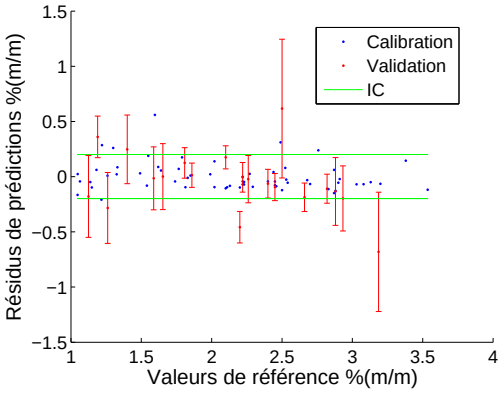
(a) Validation croisée pour le choix du nombre de facteurs PLS



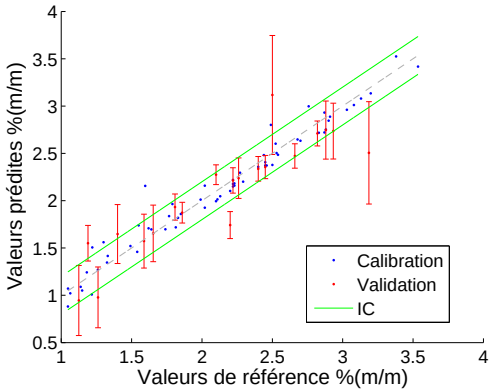
(b) Pourcentage cumulé de variance expliquée sur **X** et sur **y** en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction



(e) Diagramme de parité

FIGURE 5.10 – Interprétation du modèle de prédiction de la teneur en soufre

5.5 Conclusions

Nous avons montré dans ce chapitre que les performances des modèles permettaient d'envisager une analyse par spectroscopie PIR des propriétés suivantes : la densité, les teneurs en saturés, en aromatiques, en résines, en asphaltènes C7, en carbone Conradson, en carbones insaturés et en hydrogène. Un gain de 12 heures par échantillon sera engendré par l'analyse de ces propriétés par spectroscopie PIR. De plus, le suivi des procédés de valorisation des produits lourds sera amélioré car certaines de ces analyses n'étaient pas disponibles en suivi pour des raisons de coûts et de délais. Le transfert de ces modèles au laboratoire est en cours. Ce travail doit néanmoins être complété par le développement d'une analyse qualitative pour l'identification des échantillons qui ne contiennent pas d'asphaltènes.

Les étalonnages multivariés développés pour la détermination des teneurs en soufre et en azote ont également été décrits. Nous avons néanmoins évoqué que les performances de ces modèles ne semblent pas suffisantes pour remplacer les méthodes de référence. Ils pourraient cependant être utilisés pour donner des indications sur l'évolution de ces teneurs.

Enfin, les performances des modèles de prédiction de la viscosité et des teneurs en carbone et en asphaltènes (SARA) ne sont pas satisfaisantes (Annexe B.2). En ce qui concerne les teneurs en carbone et en asphaltènes (SARA), le faible nombre d'échantillons disponibles semble être la principale cause de non-performance. Une collecte de nouveaux échantillons devrait donc potentiellement améliorer la qualité des prédictions. Pour la viscosité, la forte non-linéarité détectée nous semble difficile à corriger. En effet, plusieurs méthodes ont été testées pour pallier ce problème mais aucune n'a été concluante. L'utilisation de méthodes de régression non-linéaires peut cependant être envisagée.

Conclusions et Perspectives

Les travaux présentés ont permis de répondre en grande partie à l'objectif général de la thèse qui était de développer une analyse rapide pour la caractérisation des produits lourds du pétrole. Ce développement permet de pallier le fait que le suivi des procédés de valorisation n'est pas optimal, le nombre d'analyses étant restreint à cause des coûts et des délais des méthodes de référence.

Le développement d'une analyse multivariée reposant sur la spectroscopie infrarouge s'est imposé comme une technique de choix pour répondre aux besoins analytiques. En effet, cette technique présente de nombreux avantages. Tout d'abord, l'acquisition de spectres est très rapide, peu coûteuse et nécessite peu de préparation des échantillons. En outre, les spectres obtenus contiennent une information très riche pour la caractérisation des hydrocarbures. Enfin, le développement de modèles d'étalonnage multivarié permet *in fine* de prédire simultanément l'ensemble des propriétés d'intérêts sur la base d'un spectre unique.

Dans la littérature, les analyses multivariées développées à partir de spectres de vibration s'appliquent de manière générale au suivi de procédés de raffinage industriels. Le but est alors la transformation du produit pour qu'il réponde aux spécifications techniques et environnementales. Dans le cadre de ce travail, l'objectif était plutôt d'analyser des produits qui sont obtenus lors de l'optimisation de plusieurs procédés de valorisation des produits lourds. Ces études ont pour but de tester de nouveaux catalyseurs et d'affiner les conditions expérimentales. La particularité de ce travail réside alors dans le

fait que les échantillons à analyser sont très disparates en termes de coupes, d'origines géographiques et de procédés. De plus, il en résulte que l'analyse multivariée doit couvrir des gammes analytiques très étendues. La première partie de ce travail de thèse a donc concerné la constitution d'une base d'échantillons représentative. La démarche pour la sélection des échantillons a tout d'abord été exposée. Nous avons ensuite démontré que la base, composée de 230 échantillons de distillats sous-vide et de résidus atmosphériques, est représentative des propriétés d'intérêt au regard des gammes analytiques considérées. De plus, nous avons montré la diversité des échantillons en termes d'origines géographiques et de procédés de valorisation.

Le bilan réalisé sur les travaux de la littérature a permis d'orienter nos choix vers les spectroscopies MIR et PIR. La viscosité et l'opacité des produits lourds nécessitent un échantillonnage spécifique et des trajets optiques courts. Le choix d'un spectromètre PIR en mode transmission, avec un trajet optique de 500 μm, a donc été fait. Dans le domaine MIR, seul le mode ATR peut répondre à une analyse rapide des produits lourds. L'acquisition des spectres MIR et PIR des produits lourds nécessite néanmoins de travailler en température. Les protocoles expérimentaux ont donc été optimisés pour limiter les variations de température et ainsi assurer une bonne fidélité de la mesure spectrale. Ces protocoles ont ensuite été systématisés pour permettre l'acquisition de l'ensemble des spectres de la base.

Lors de l'interprétation des spectres des produits lourds, nous avons constaté, comme escompté, que la plupart des bandes sont dues aux vibrations de liaisons C-H et C=C. Une courbure de la ligne de base a néanmoins été observée sur certains spectres PIR sur le domaine 9000-6000 cm^{-1}. En nous appuyant sur la littérature, cet effet a été attribué à une absorption électronique des liaisons $n-\pi^*$ et $\pi-\pi^*$ causée par les agrégats d'asphaltènes.

L'analyse exploratoire a permis de dévoiler que la base spectrale PIR se divise en deux lots associés à la présence ou non de cette absorption électronique. Cette distribution des spectres PIR est problématique pour le développement des modèles de prédiction des propriétés pour lesquelles l'information liée à l'absorption électronique n'est pas pertinente. Comme nous l'avons évoqué, les méthodes de prétraitements mathématiques de spectres et la sélection des variables sont les deux principales voies pour l'optimisation des modèles

en vue de corriger ce phénomène. De plus, nous avons abordé la nécessité de réduire la variabilité de l'absorption électronique pour le développement du modèle de prédiction de la teneur en asphaltènes C7. Dans le cadre de l'optimisation de ce modèle, il existe alors une forte interaction entre la méthode de prétraitement appliquée et les variables sélectionnées.

Le potentiel d'une méthode d'optimisation simultanée du choix des prétraitements et des variables à sélectionner par AG a donc été évalué. Des modèles "préliminaires", basés sur les travaux de la littérature, ont tout d'abord été développés en appliquant la dérivée 1ère. Ces modèles ont ensuite fait office de "référence" pour évaluer l'apport de la procédure d'optimisation par AG. Pour ce faire, un test statistique, le "randomisation t-test", a été appliqué pour la comparaison des modèles. L'approche d'optimisation proposée a apporté des réponses sur le choix des méthodes de correction des spectres lors de l'optimisation du prétraitement seul, sur un domaine spectral continu. D'une part, la dérivée 1ère a été sélectionnée lors de l'optimisation par AG pour la détermination des teneurs en asphaltènes C7. Cette méthode est donc le meilleur compromis pour réduire la variabilité associée aux contributions de l'absorption électronique tout en conservant l'information qu'elle contient. D'autre part, il a été établi que la méthode *Weighted Least Square Baseline* (WLSB) avec un polynôme d'ordre 3 est la technique la plus appropriée pour éliminer les variations de l'absorption électronique lorsqu'elle n'est pas une information utile pour la description de la propriété considérée. La procédure par algorithmes génétiques a également montré son potentiel dans le cadre de l'optimisation simultanée du choix des prétraitements et des variables à sélectionner. En effet, nous avons démontré un apport significatif de la co-optimisation pour la détermination de la teneur en aromatiques. Enfin, les variables sélectionnées peuvent être interprétées et sont globalement en accord avec la composition chimique des fractions SARA.

Nous avons également procédé à une étude de comparaison et de fusion des spectroscopies MIR et PIR pour le développement de l'analyse multivariée. Cette étude nous a tout d'abord permis de définir que la spectroscopie PIR est globalement plus performante, dans le cadre de notre application, pour la prédiction des propriétés des produits lourds. En effet, les modèles de prédiction des teneurs en saturés, aromatiques et carbones insa-

turés développés à partir des spectres PIR sont significativement plus performants que ceux développés à partir de la spectroscopie MIR. Pour les autres propriétés, les modèles développés à partir de ces deux techniques spectroscopiques sont équivalents.

L'étude sur l'exploitation simultanée des spectroscopies MIR et PIR nous a ensuite permis d'apporter des réponses sur le potentiel de la fusion de données en termes de capacités de prédiction. Nous avons constaté que, pour espérer une amélioration globale du pouvoir prédictif lors de la fusion de données spectroscopiques, il est nécessaire que ces deux techniques aient des performances similaires lorsqu'elles sont considérées séparément. En effet, les erreurs de RMSEP des modèles MB-PLS et S-PLS pour la détermination des teneurs en résines et en asphaltènes C7 sont plus faibles que celles des modèles développés en considérant séparément les spectres MIR et PIR. Une amélioration significative a notamment pu être démontrée par le modèle MB-PLS de prédiction des teneurs en résines. En revanche, pour les teneurs en saturés et en aromatiques, la fusion de données améliorent le pouvoir prédictif comparée à celui des modèles calculés sur les spectres MIR. Aucune amélioration significative n'a pu être démontrée par rapport aux modèles développés à partir de la spectroscopie PIR. Les méthodes MB-PLS et S-PLS ont également confirmé leur potentiel d'interprétation. En effet, nous avons pu mettre en évidence que l'amélioration des prédictions des teneurs en résines lors de la fusion de données sont dues à la contribution de trois bandes fondamentales : la déformation des liaisons C-H hors du plan, l'élongation des doubles liaisons C=C et, dans une moindre mesure, l'élongation des liaisons =C-H.

La dernière partie de ce manuscrit a présenté les modèles de prédiction des autres propriétés d'intérêt pour la caractérisation globale des produits lourds du pétrole. Dans ce cadre, nous avons également appliqué une technique de *bootstrap* pour l'encadrement des valeurs prédites. Nous tenons à rappeler que ces modèles ont été obtenus assez tôt dans le travail de thèse et n'ont pas fait l'objet d'une optimisation spécifique approfondie. Les étalonnages multivariés de huit propriétés ont été jugés satisfaisants en termes d'erreur de prédiction. Ces propriétés sont la densité et les teneurs en saturés, en aromatiques, en résines, en asphaltènes C7, en carbone Conradson, en carbones insaturés et en hydrogène. Pour les teneurs en azote et en soufre, les corrélations obtenues sont relativement satisfai-

santes mais ne peuvent, en l'état, se substituer aux analyses de référence. Pour conclure, l'application en laboratoire de cette analyse rapide des produits lourds va engendrer un gain de temps d'analyse d'environ 12 heures. De plus, l'analyse spectroscopique va permettre de proposer des analyses pour le suivi des procédés qui ne sont pas disponibles à présent.

En ce qui concerne la suite de ce travail, le transfert de ces modèles au laboratoire pour une utilisation en routine est en cours. Une analyse qualitative doit néanmoins être développée pour identifier les échantillons dont les teneurs en asphaltènes sont nulles. Certaines propriétés n'ont pas pu être déterminées en l'état par spectroscopie PIR. En effet, le nombre d'échantillons disponibles est insuffisant pour obtenir un étalonnage multivarié satisfaisant pour les teneurs en carbone et en asphaltènes. Une collecte d'échantillons est donc à prévoir. De plus, la non-linéarité entre les spectres PIR et la viscosité n'a pas pu être corrigée par les approches que nous avons testées. Une évaluation de méthodes de régression capables de pallier les non-linéarités peut alors être envisagée.

Des implémentations de la procédure d'optimisation par algorithmes génétiques ont également été proposées. D'une part, une restriction sur le choix des prétraitements lors de l'optimisation peut être mise en place afin d'éviter des combinaisons de méthodes aberrantes et ainsi, d'évaluer l'apport de la combinaison de trois ou quatre prétraitements. L'implémentation d'une méthode telle que la Bi-PLS est également envisagée dans le but de s'assurer de l'apport des variables sélectionnées à la description de la propriété. Une interprétation plus fidèle des variables sélectionnées pourrait être obtenue par l'application de cette méthode au modèle final.

Afin d'étendre l'analyse développée au cours de ce travail à l'ensemble des besoins analytiques des procédés de valorisation des produits lourds, une analyse rapide des résidus sous-vide par spectroscopie MIR en mode ATR a fait l'objet d'une première étude de faisabilité qui s'est avérée concluante. De plus, une étude de faisabilité de la prédiction des propriétés de sous-coupes d'un produit est également envisagée. Cette analyse consisterait, par exemple, à prédire les valeurs des propriétés du distillat sous-vide (350-550°C) et du résidu sous-vide (550°C$^+$) à partir de l'analyse spectroscopique du résidu atmosphérique (350°C$^+$). Dans ce cas, cette application permettrait de s'affranchir de la distillation sous-

vide qui est longue et qui nécessite d'importants volumes d'échantillons.

Bibliographie

[1] *BP Statistical Review*, 2011.

[2] C Abrahamsson, J Johansson, A Sparen et F Lindgren : Comparison of different variable selection methods conducted on nir transmission measurements on intact tablets. *Chemometrics and Intelligent Laboratory Systems*, 69:3–12, 2003.

[3] M.C.U Araújo, T.C.B Saldanha, R.K.H Galvão, T Yoneyama, H.C Chame et Visani : The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2): 65–73, 2001.

[4] N Aske, H Kallevik et J Sjöblom : Determination of saturate, aromatic, resin and asphaltenic (sara) components in crude oils by means of infrared and near-infrared spectroscopy. *Energy & Fuels*, 15:1304–1312, 2001.

[5] R.J Barnes, M.S Dhanoa et S.J Lister : Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5):737–891, 1989.

[6] K Baumann, H Albert et M Von Korff : A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. part ii. pratical applications. *Journal of Chemometrics*, 16:339–350, 2002.

[7] D Beasley, D.R Bull et R.R Martin : An overview of genetic algorithms : Part 1, fundamentals ". *University Computing*, 15:58–59, 1993.

[8] D Beasley, D.R Bull et R.R Martin : An overview of genetic algorithms : Part 2, research topics. *University Computing*, 15:170–181, 1993.

[9] K.R Beebe et B.R Kowalski : An introduction to multivariate calibration and analysis. *Analytical Chemistry*, 59(17):A1007, 1987.

[10] J Beens et U.A.T Brinkman : The role of gas chromatography in compositional analyses in the petroleum industry. *TrAC, Trends in Analytical Chemistry*, 19:260–275, 2000.

[11] A Berglund et S Wold : A serial extension of multiblock pls. *Journal of Chemometrics*, 13:461–471, 1999.

[12] D Bertrand et E Dufour : *La spectroscopie infrarouge*, volume 2. Lavoisier, 2006.

[13] M Blanco, S Maspoch, I Villarroya, X Peralta, J.M Gonzalez et J Torres : Determination of the penetration value of bitumens by near infrared spectroscopy. *Analyst*, 125:1823–1828, 2000.

[14] M Blanco, S Maspoch, I Villarroya, X Peralta, J.M Gonzalez et J Torres : Determination of physical properties of bitumens by use of near-infrared spectroscopy with neural networks. joint modelling of linear and non-linear parameters. *Analyst*, 126(378-382), 2001.

[15] M Blanco, S Maspoch, I Villarroya, X Peralta, J.M Gonzalez et J Torres : Determination of physico-chemical parameters for bitumens using near infrared spectroscopy. *Analytica Chimica Acta*, 434:133–141, 2001.

[16] Z Boger : Selection of quasi-optimal inputs in chemometrics modeling by artificial neural network analysis. *Analytica Chimica Acta*, 490(1-2):31–40, 2003.

[17] C.W Brown, P.F Lynch, R.J Obremski et D.S Lavery : Matrix representations and criteria for selecting analytical wavelengths for multicomponent spectroscopic analysis. *Analytical Chemistry*, 54(9):1472–1479, 1982.

[18] PB Bràs, J.A Bernardino, S.A Lopes et J.C Menezes : Multiblock pls as an approach to compare and combine nir and mir spectra in calibrations of soybean flour. *Chemometrics and Intelligent Laboratory Systems*, 75:91–99, 2005.

[19] D.A Burns et E.W Ciurczak : *Handbook of near-infrared spectroscopy*, volume 13. Marcel Dekker, Inc, 1992.

[20] V Centner, D.L Massart, O.E deNoord, S de Jong, B.M Vandeginste et Sterna C : Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry*, 68(21):3851–3858, 1996.

[21] V Centner, L Verdu-Andres, B Walczak, D Jouan-Rimbaud, F Despagne, L Pasti, R Poppi, D.L Massart et O.E de Noord : Comparison of multivariate calibration techniques applied to experimental nir data sets. *Applied Spectroscopy*, 54(4):608–622, 2000.

[22] D Chen, W Cai et X Shao : Representative subset selection in modified iterative predictor weighting (mipw) - pls models for parsimonious multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 87(2):312–318, 2007.

[23] H Chung : Applications of near-infrared spectroscopy in refineries and important issues to address. *Applied Spectroscopy Reviews*, 42:251–285, 2007.

[24] H Chung et M.S Ku : Comparison of near-infrared, infrared and raman spectroscopy for the analysis of heavy petroleum products. *Applied Spectroscopy*, 54(2):239–245, 2000.

[25] H Chung et M.S Ku : Near-infrared spectroscopy for on-line monitoring of lube base oil process. *Applied Spectroscopy*, 57(5):545–550, 2003.

[26] G Cruciani, M Baroni, S Clementi, G Costantino, D Riganelli et B Skagerberg : Predictive ability of regression-models. 1. standard-deviation of prediction errors (sdep). *Journal of Chemometrics*, 6:335–346, 1992.

[27] P.J De Groot, G.J Postma, Melssen W.J et Buydens L.M.C : Selecting a representative training set for the classification of demolition waste using remote nir sensing. *Analytica Chimica Acta*, 392(1):67–75, 1999.

[28] L.F.B de Lira, M.S de Albuquerque, J.G.A Pacheco, T.M Fonseca, E.H.D Cavalcanti, L Stragevitch et M.F Pimentel : Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel. *Microchemical journal*, 96(1):126–131, 2010.

[29] P De Peinder, T Visser, D.D Petrauskas, F Salvatori, F Soulimani et B.M Weckhuysen : Partial least squares modeling of combined infrared, ^{1}h nmr and ^{13}c nmr spectra to predict long residue properties of crude oils. *Vibrational Spectroscopy*, 51:205–212, 2009.

[30] D.M Deaven et K.M Ho : Molecular-geometry optimization with a genetic algorithm. *Physical review letters*, 75(2):288–291, 1995.

[31] O Devos et Duponchel L : Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for pls regression. *Chemometrics and Intelligent Laboratory Systems*, 107(1):50–58, 2011.

[32] A Doucet, N De Freitas et N Gordon : *Sequential Monte Carlo methods in practice*. Springer, New York, 2001.

[33] Y.P Du, Y.Z Liang, J.H Jiang, R.J Berry et Y Ozaki : Spectral regions selection to improve prediction ability of pls models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica Chimica Acta*, 501(2):183–191, 2004.

[34] A Durand : *Méthodes de sélection de variables appliquées en spectroscopie proche infrarouge pour l'analyse et la classification de textiles*. Thèse de doctorat, Université des sciences et technologies de Lille - Ecole doctorale des sciences pour l'ingénieur, 2007.

[35] A Durand, O Devos, C Ruckebusch et J-P Huvenne : Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton-viscose textiles. *Analytica Chimica Acta*, 595:72–79, 2007.

[36] B Efron : Bootstrap methods : Another look at the jackknife. *Annals of statistics*, 7(1):1–26, 1979.

[37] B Efron et R.J Tibshirani : *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[38] P Eilers et B Marx : Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:89–121, 1996.

[39] J Eyssautier, P Levitz, D Espinat, J Jestin, J Gummel, I Grillo et L Barre : Insight into asphaltene nanoaggregate structure inferred by small angle neutron and x-ray scattering. *Journal of physical chemistry B*, 115:6827–6837, 2011.

[40] M Feinberg : *L'assurance qualité dans les laboratoires agroalimentaires et pharmaceutiques*. Tech. & Doc., Lavoisier, 2ème édition, 2001.

[41] C.C Felicio, L.P Bràs, J.A Lopes, L Cabrita et J.C Menezes : Comparison of pls algorithms in gasoline and monitoring with mir and nir. *Chemometrics and Intelligent Laboratory System*, 78(1-2):74–80, 2005.

[42] R.K.H Galvão, M.C.U Araujo, G.E José, M.J.C Pontes, E.C da Silva et T.C.B Saldanha : A method for calibration and validation subset partitioning. *Talanta*, 67:736–740, 2005.

[43] P Geladi et B.R Kowalski : Partial least-squares regression - a tutorial. *Analytica Chimica Acta*, 185:19–32, 1986.

[44] P Geladi, D MacDougal et H Martens : Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3):491–500, 1985.

[45] W.R Gilbert, F.S Gusmao de Lima et A.F Bueno : Comparison of nir and nmr spectra chemometrics for fcc feed online characterization. *Studies in Surface Science and Catalysis*, 149, 2004.

[46] G.F Giskeodegard, M.T Grinde, B Sitter, D.E Axelson, S Lundgren, HE Fjosne, S Dahl, I.S Gribbestad et TF Bathen : Multivariate modeling and prediction of breast cancer prognostic factors using mr metabolomics. *Journal of protrome research*, 9(2):972–979, 2010.

[47] D.E Goldberg : *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading MA, Addison-Wesley, 1989.

[48] C.M Grinstead et J.L Snell : *Introduction to probability*. American Mathematical Society, 1997.

[49] D.M Haaland et E.V Thomas : Partial least-squares methods for spectral analyses. 1 .relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, 60(11):1193–1202, 1988.

[50] L Hadjiiski, P Geladi et P Hopke : A comparison of modeling nonlinear systems with artificial neural networks and partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 49:91–103, 1999.

[51] A Hannisdal, P.I.V Hemmingsen et J Sjöblom : Group-type analysis of heavy crude oils using vibrational spectroscopy in combination with multivariate analysis. *Industrial and Engineering Chemistry Research*, 44:1349–1357, 2005.

[52] K Hidajat et S.M Chong : Quality characterisation of crude oils by partial least square calibration of nir spectral profiles. *Journal of Near Infrared Spectroscopy*, 8:53–59, 2000.

[53] J.H Holland : *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence*. MIT Press, 1992.

[54] Y Hongfu, C Xiaoli, L Hoaran et X Yupeng : Determination of multi-properties of residual oils using mid-infrared attenuated total reflection spectroscopy. *Fuel*, 85:1720–1728, 2006.

[55] L Ingber : Very fast simulated re-annealing. *Mathematical and computer modelling*, 12(8):967–973, 1989.

[56] D Jouan-Rimbaud, D.L Massart, R Leardi et O.E De Noord : Genetic algorithms as a tool for wavelength selection in multivariate calibration. *Analytical Chemistry*, 67(23):4295–4301, 1995.

[57] D Jouan-Rimbaud, D.L Massart, C.A Saby et C Puel : Characterisation of the representativity of selected sets of samples in multivariate calibration and pattern recognition. *Analytica Chimica Acta*, 350(1-2):149–161, 1997.

[58] J.H Kalivas, N Roberts et J.M Sutter : Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry. *Analytical Chemistry*, 61(18):2024–2030, 1989.

[59] S Kasemsumran, Y.P Du, K Maruo et Y Ozaki : Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 82:97–103, 2006.

[60] R.W Kennard et L.A Stone : Computer aided design of experiments. *Technometrics*, 11:137–148, 1969.

[61] S Kirkpatrick et M.P Vecchi : Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[62] J Koljonen, T.E.M Nordling et J.T Alander : A review of genetic algorithms in near infrared spectroscopy and chemometrics : past and future. *Journal of Near Infrared Spectrosopy*, 16(3):189–197, 2008.

[63] T Kourti, P Nomikos et J.F MacGregor : Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls. *Journal of process control*, 5(4):277–284, 1995.

[64] R Leardi : Genetic algorithms in chemometrics and chemistry : a review. *Journal of Chemometrics*, 15:559–569, 2001.

[65] R Leardi : Genetic algorithms in chemistry. *Journal of Chromatography A*, 1158: 226–233, 2007.

[66] R Leardi, R Boggia et M Terrile : Genetic algoritms as a strategy for feature selection. *Journal of chemometrics*, 6:267–281, 1992.

[67] R Leardi et M Lupiáñez : Genetic algorithms applied to feature selection in pls regression : how and when to use them. *Chemometrics and intelligent laboratory systems*, 41(5-6):195–207, 1998.

[68] R Leardi et L Nørgaard : Sequential application of backward interval partial least squares and genetic algorithms for the selection of revelant spectral regions. *Journal of chemometrics*, 18:486–497, 2004.

[69] F.S.G Lima, M.A.S Araùjo et L.E.P Borges : Determination of lubricant base oil properties by near infrered spectrosopy using different sample and variable selection methods. *Journal of Near Infrared Spectroscopy*, 12:159–166, 2004.

[70] F.S.G Lima et L.F.M Leite : Determination of asphalt cement properties by near infrared spectroscopy and chemometrics. *Journal of Petroleum Science and Engineering*, 22(5 and 6):589–600, 2004.

[71] W Lindberg, J-A Persson et S Wold : Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and lignin sulfonate. *Analytical Chemistry*, 55:643–648, 1983.

[72] Y Long, T Dabros et H Hamza : Analysis of solvent-diluted bitumen from oil sands froth treatment using nir spectroscopy. *The Canadian Journal of Chemical Engineering*, 82:776–781, 2004.

[73] J.A Lopes, J.C Menezes, J.A. Westerhuis et A.K Smilde : Multiblock pls analysis of an industrial pharmaceutical process. *Biotechnology and bioengineering*, 80(4):419–427, 2002.

[74] JF Macgregor, Jaeckle, C Kiparissides et M Koutoudi : Process monitoring and diagnosis by multiblock pls methods. *Aiche Journal*, 40(5):826–838, 1994.

[75] H Mark : Comparative-study of calibration methods for near-infrared reflectance analysis using a nestes experimental-design. *Analytical Chemistry*, 58(13):2814–2819, 1986.

[76] H Martens, S.A Jensen et P Geladi : Multivariate linearity transformations for near infrared reflectance spectroscopy. *Applied Statistics*, pages 205–234, 1983.

[77] H Martens et T. Naes : *Multivariate Calibration*. John Wiley & Sons, 1989.

[78] A.M McKenna, J.M Purcell, R.P Rodgers et A.G Marshall : 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 & Fuels*, 24:2929–2938, 2010.

[79] I Merdrignac et D Espinat : Physicochemical characterization of petroleum fractions : the state of the art. *Oil & Gas Science and Technology - Rev.IFP*, 62(1):7–32, 2007.

[80] O.C Mullins : Asphaltenes in crude oil : absorbers and/or scatterers in the near-infrared region ? *Analytical Chemistry*, 62(5):508–514, 1990.

[81] J Murgich, J.M Rodriguez et Y Aray : Molecular recognition and molecular mechanics of micelles of some model asphaltenes and resins. *Energy & Fuels*, 10:68–76, 1996.

[82] K.E Nielsen, J Dittmer, A Malmendal et N.Chr Nielsen : Quantitative analysis of constituents in heavy fuel oil by 1h nuclear magnetic resonance (nmr) spectroscopy and multivariate data analysis. *Energy & Fuels*, 22:4070–4076, 2008.

[83] A.S.L Nørgaard, J Wagner, J.P Nielsen, L Munck et S.B Engelsen : Interval partial least-squares regression (ipls) : A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54(3):413–419, 2000.

[84] B.G Osborne et T Fearn : *Near Infrared spectroscopy in food analysis*. Longman Scientific & Technical, 1986.

[85] B.G Osborne, T Fearn et P.H Hindle : *Practical NIR spectroscopy with applications in food and beverage analysis*. Harlow : Prentice Hall Ed, 1993.

[86] C Pasquini et A.F Bueno : Characterization of petroleum using near-infrared spectroscopy : Quantitative modeling for the true boiling point curve and specific gravity. *Fuel*, 86:1927–1934, 2007.

[87] R Poilblanc et F Crasnier : *Spectroscopies infrarouge et Raman*. EDP Sciences, 2006.

[88] A Rinnan, F Van Den Berg et S Balling Engelsen : Review of the most common pre-processing techniques for near-infrared spectra. *Analytical Chemistry*, 28(10):1201–1222, 2009.

[89] L Riveros, B Jaimes, M.A Ranaudo, J Castillo et J Chirinos : Determination of asphaltene and resin content in venezuelan crude oils by using fluorescence spectroscopy and partial least squares regression. *Energy & Fuels*, 20:227–230, 2006.

[90] Y Roggo, L Duponchel, C Ruckebusch et J-P Huvenne : Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data. *Journal of molecular structure*, 654:253–262, 2003.

[91] B Rosner et M.L.T Glynn, R.J Lee : The wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, 62:185–192, 2006.

[92] C Ruckebusch, F Orhan, A Durand, T Boubellouta et J-P Huvenne : Quantitative analysis of cotton-polyester textile blends from near-infrared spectra. *Applied Spectroscopy*, 60(5):120A–148A, 2006.

[93] S Satya, R.M Roehner, M.D Deo et F.V Hanson : Estimation of properties of crude oil residual fractions using chemometrics. *Energy & Fuels*, 2007.

[94] A Savitsky et M.J.E Golay : Smoothing and differentiation of data by simplified least-squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[95] J Snyman : *Practical Mathematical Optimization : An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*. Springer, New York, 2005.

[96] J.G Speight : Petroleum asphaltenes - part 1 - asphaltenes, resins and the structure of petroleum. *Oil & Gas Science and Technology - Rev.IFP*, 59(5):467–477, 2004.

[97] C.H Spiegelman, McShane M.J, Goetz M.J, M Motamedi, Q.L Yue et G.L Cote : Theoretical justification of wavelength selection in pls calibration : Development of a new algorithm. *Analytical Chemistry*, 70:35–44, 1998.

[98] H Swierenga, F Wülfert, O.E de Noord, A.P de Weijer, A.K Smilde et L.M.C Buydens : Development of robust calibration models in near infra-red spectrometric applications. *Analytica Chimica Acta*, 411(1-2):121–135, 2000.

[99] H.H Szu et Hartley R.L : Fast simulated annealing. *Physics letters A*, 122(3-4):157–162, 1987.

[100] M Tenenhaus : *La régression PLS : Théorie et Pratique*. Editions technip, 1998.

[101] M.L Thompson : Selection of variable in multiple regression, part 1 : A review and evaluation. *International Statistical Review*, 46:1–19, 1978.

[102] R Todeschini, D Galvagni, J.L Vilchez, M del Olmo et N Navas : Kohonen artificial neural networks as a tool for wavelength selection in multicomponent spectrofluorimetric pls modelling : application to phenol, o-cresol, m-cresol and p-cresol mixtures. *TrAC, Trends in Analytical Chemistry*, 18(2):93–98, 1999.

[103] H Van Der Voet : Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory System*, 25:313–323, 1994.

[104] R.Z Wang, C.F Lin et J.C Lin : Image hiding by optimal lsb substitution and genetic algorithm. *Pattern Recognition*, 34(3):671–683, 2001.

[105] L.E. Wangen et B.R Kowalski : A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, 3:3–20, 1988.

[106] J-P Wauquier : *Le raffinnage du pétrole*, volume 1. Editions Technip, 1994.

[107] J.A Westerhuis et P.M.J Coenegracht : Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of chemometrics*, 11:379–392, 1997.

[108] J.A Westerhuis, T Kourti et J.F Macgregor : Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12:301–321, 1998.

[109] J.A Westerhuis et A.K Smilde : Deflation in multiblock pls. *Journal of Chemometrics*, 15:485–493, 2001.

[110] S Wold, S Hellberg, T Lundstedt, M Sjostrom et H Wold : Proceedings symposium on pls model building : Theory and application. 1987.

[111] S Wold, N Kettaneh et K Tjessem : Hierarchical multiblock pls and pc models for easier model interpretation and as an alternative to variable selection. *Journal of chemometrics*, 10:463–482, 1996.

[112] S Wold, M Sjostrom et L Eriksson : Pls-regression : a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[113] Z Xiabo, Z Jiewen, M.J.W Povey, M Holmes et M Hanpin : Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, 667:14–32, 2010.

[114] T.F Yen, J.G Erdman et S.S Pollack : Investigation of the structure of petroleum asphaltenes by x-ray diffraction. *Analytical Chemistry*, 33:1587–1594, 1961.

[115] Y. Zheng, X Lai, S.W Bruun, H Ipsen, J. N Larsen, H Løwenstein, I Søndergaard et S Jacobsen : Determination of moisture content of lyophilized

allergen vaccines by nir spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 46(3):592–596, 2008.

[116] X Zhu, S Yang, G Li, A Huang et Z Zhang : Prediction of wood property in chinese fir based on visible/near-infrared spectroscopy and least square-support vector machine. *Spectrochimica Acta Part A*, 74:344–348, 2009.

A Les techniques chimiométriques complémentaires

Dans cette annexe, les méthodes couramment appliquées en chimiométrie et exploitées durant cette thèse seront présentées. Nous décriront tout d'abord les techniques de prétraitement mathématiques des spectres. Les méthodes pour la sélection des échantillons d'étalonnage et de validation seront ensuite exposées. La régression PLS et la validation croisée seront ensuite détaillées. Enfin, les différents critères statistiques pour l'évaluation des performances des modèles sont rappelés.

A.1 Prétraitements mathématiques

Les spectres de vibrations peuvent être affectés par des interférences physico-chimiques qui ne sont pas reliées à la composition du produit analysé : du bruit (une erreur non systématique), de la diffusion, des variations du trajet optique, une non-linéarité du détecteur dans certaines gammes d'absorbance...

De ce fait, il est nécessaire d'appliquer des prétraitements mathématiques aux spectres. L'objectif d'un prétraitement de spectres est de réduire les interférences physico-chimiques présentes dans les spectres tout en maintenant la variabilité due à la composition chimique des échantillons dans le but d'améliorer les modèles [88].

A.1.1 Dérivation

Les méthodes de dérivation ont la capacité de corriger à la fois les effets additifs (déplacements verticaux de la ligne de base ou "offset") et multiplicatifs (déplacements

verticaux de la ligne de base en fonction de la longueur d'onde) qui peuvent apparaître dans les spectres [88]. La figure A.1 montre l'effet de la dérivée sur un spectre affecté par un effet additif (vert) et par un effet multiplicatif (rouge). La dérivée 1$^{\text{ère}}$ est capable de corriger l'un des effets (additif pour le spectre vert et multiplicatif pour le spectre rouge). Cependant, lorsque les deux phénomènes sont en présence, l'utilisation de la dérivée 2$^{\text{nde}}$ est nécessaire.

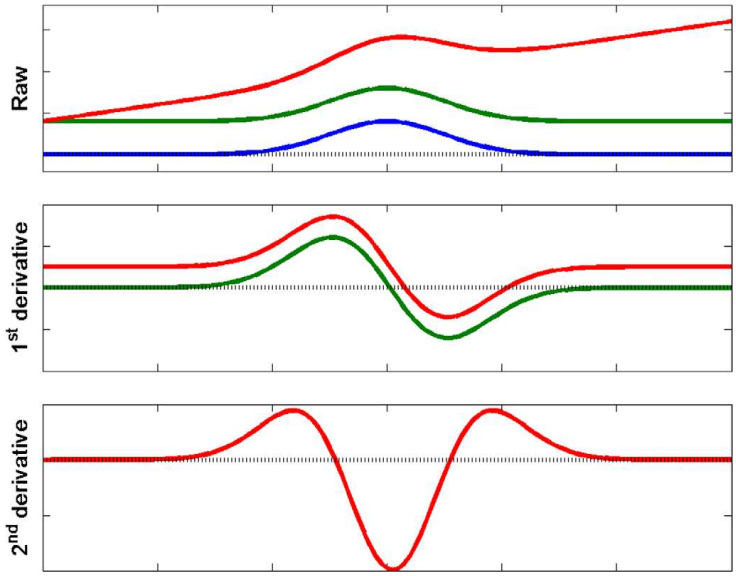


FIGURE A.1 – Effet de la dérivée sur les effets additifs (vert) et additifs plus multiplicatifs (rouge). Le spectre bleu représente le spectre sans effet d'offset et la ligne noire est l'axe des abscisses [88]

La méthode la plus couramment utilisée en chimiométrie pour la dérivation des spectres est l'algorithme de SAVITZKY-GOLAY [94]. Les méthodes de dérivation entraînent généralement une diminution du rapport signal sur bruit ce qui affecte la qualité spectrale et donc les performances des modèles. Afin de réduire ce phénomène, l'algorithme de SAVITZKY-GOLAY utilise une technique de lissage. En effet, cette méthode calcule la dérivée en chaque point (longueur d'onde) i du spectre en deux étapes. Premièrement, un polynôme de degré k (généralement k = 2 ou 3) est ajusté autour du point i sur f points du spectre (avec f $\geq$ k+1). La dérivée d'ordre m (généralement 1 ou 2) du polynôme en ce point i est ensuite calculée. L'algorithme de SAVITZKY-GOLAY peut également être utilisé pour le lissage en fixant l'ordre de la dérivée à zéro.

A.1.2 Méthodes de correction de ligne de base

Les méthodes de correction de ligne de base sont généralement utilisées pour corriger les courbures qui peuvent apparaître dans le spectre. Ces courbures correspondent à une augmentation des valeurs d'absorbance, généralement dues aux effets de diffusion.

A.1.2.1 Méthode "Detrend"

La méthode "Detrend" consiste à ajuster un polynôme de degré K au spectre puis à le soustraire au spectre initial [5]. Cette méthode peut donc s'écrire, pour chaque point i du spectre, sous la forme : $x_{i,detrend} = x_{i,initial} - d_i$ où $x_{detrend}$ représente le spectre prétraité, $x_{initial}$ le spectre initial et d le polynôme ajusté par la méthode "detrend".

A.1.2.2 Méthode *Weighted Least Square Baseline* (WLSB)

La méthode "Weighted Least Square Baseline" consiste également à ajuster un polynôme de degré K au spectre puis à le soustraire au spectre initial [38]. Cependant, cette méthode utilise un algorithme itératif basé sur les moindres carrés et affecte un poids à chaque point du spectre selon le principe suivant :

- Si $x_{i,initial} - x_{wlsb} > 0$: un poids faible est attribué au point i
- Si $x_{i,initial} - x_{wlsb} < 0$: un poids fort est attribué au point i

En effet, lorsque le résidu $(x_{i,initial} - x_{wlsb})$ est négatif, le polynôme ajusté se trouve "au dessus" du spectre. Or, l'objectif est d'ajuster un polynôme sur la ligne de base. Un poids fort est alors attribué à ces points afin de "forcer" le polynôme à s'ajuster "au dessous" du spectre. Lorsque les résidus sont positifs, le raisonnement inverse est effectué.

A.1.3 Normalisation

Les méthodes de normalisation ont pour but de réduire les variations entre les échantillons dues par exemple à la diffusion et de corriger les déplacements verticaux de la ligne de base. Les corrections *Multiplicative Scatter Correction* (MSC) et *Standard normal Variate* (SNV) sont méthodes les plus couramment utilisées.

A.1.3.1 La correction "Multiplicative Scatter Correction" (MSC)

Le prétraitement MSC a pour but de corriger les problèmes de diffusion ou de variation du trajet optique [44, 76]. Le principe est de corriger chaque spectre sur la base d'un spectre de référence qui est souvent le spectre moyen. Un modèle linéaire est tout d'abord ajusté entre le spectre x_i et le spectre moyen x_m selon l'équation A.1.

$$x_i = a_i + b_i x_m + e_i \tag{A.1}$$

avec a et b, les coefficients de la régression calculés pour chaque spectre x_i. Le spectre est ensuite corrigé au moyen de l'équation A.2.

$$x_{i,corr} = \frac{x_i - a_i}{b_i} \tag{A.2}$$

où $x_{i,corr}$ est le spectre obtenu par l'application de cette correction MSC.

A.1.3.2 La correction "Standard Normal Variate" (SNV)

La correction SNV [5] a pour but de corriger les effets de déplacements verticaux de la ligne de base. Elle est basée sur le calcul de l'écart type de l'absorbance à chaque longueur d'onde du spectre et s'applique à chaque spectre pris séparément, sans référence à l'ensemble des échantillons. Cette correction se calcule au moyen de l'équation :

$$SNV_i = \frac{y_i - \overline{y}}{\sqrt{\frac{\sum (y_i - \overline{y})^2}{n-1}}} = \frac{y_i - \overline{y}}{\sigma} \tag{A.3}$$

où $\overline{y}$ est la moyenne des absorbances du spectre, y_i l'absorbance à corriger, n le nombre de longueurs d'onde du spectre, σ l'écart type à la moyenne des absorbances du spectre et SNV_i l'absorbance corrigée. Ce prétraitement est souvent utilisé lors de l'acquisition en réflexion diffuse sur des échantillons en poudre ou comportant des particules.

A.2 Méthodes de sélection des échantillons

L'un des points importants pour le développement d'une analyse multivariée est la sélection des échantillons des lots d'étalonnage et de validation [27, 57]. En effet, les échantillons du lot d'étalonnage doivent couvrir l'espace de la base initiale et, notamment, les extrémités. Pour garantir une bonne évaluation des performances des modèles, les échantillons du lot de validation doivent couvrir le même espace de variation et se trouver dans l'espace défini par les échantillons d'étalonnage. Nous souhaitons de cette manière éviter les problèmes d'extrapolation.

De nombreuses techniques de sélection d'échantillons peuvent être utilisées. Dans cette partie, nous présenterons trois d'entre elles : la sélection aléatoire, la méthode de "Kennard et Stone" et la méthode SPXY.

A.2.1 La sélection aléatoire

La méthode de sélection aléatoire est une technique rapide et simple pour la sélection d'échantillons. Cette technique peut s'avérer efficace sur les lots importants car le tirage d'un groupe d'échantillons dans une grande population suit la distribution statistique du lot entier. Cependant, la sélection aléatoire n'assure pas la représentativité du lot complet, notamment sur les lots avec peu d'échantillons et n'empêche pas les problèmes d'extrapolation.

A.2.2 Méthode de "Kennard et Stone"

La méthode de Kennard et Stone [60] est basée sur les distances euclidiennes entre les échantillons. La distance euclidienne d_x entre deux vecteurs p et q de taille j se calcule selon l'équation A.4.

$$d_x(p,q) = \sqrt{\sum_{j=1}^{j}[x_p(j) - x_q(j)]^2} \tag{A.4}$$

Cette méthode est initialisée en sélectionnant, soit les deux échantillons les plus éloignés, soit l'échantillon le plus au centre de la base. Ensuite, à chaque itération, les distances entre les échantillons déjà sélectionnés et les échantillons restants sont calculées. L'échan-

tillon qui admet la plus grande des distances est à son tour sélectionné. Ces opérations sont réitérées jusqu'à ce que le nombre d'échantillons défini au préalable soit atteint. Les échantillons sélectionnés constituent le lot de d'étalonnage et les échantillons restants forment le lot de validation. La méthode de KENNARD et STONE peut-être appliquée sur les spectres bruts, les spectres prétraités, sur les scores ACP ou sur les scores PLS.

A.2.3 Méthode SPXY

La méthode SPXY, proposée par Galvão *et al.* [42], est similaire à la méthode de "KENNARD et STONE". La principale différence est que la méthode SPXY prend également en compte la distance euclidienne entre les valeurs de la propriété. Pour cela, une distance $d_{x,y}$ combinant les distances entre les spectres d_x et les valeurs de la propriété d_y est calculée (Equation A.5).

$$d_{x,y}(p,q) = \frac{d_x(p,q)}{max_{p,q\in[1,N]}(d_x)} + \frac{d_y(p,q)}{max_{p,q\in[1,N]}(d_y)} \tag{A.5}$$

Les distances euclidiennes d_x et d_y sont normalisées par la valeur maximale de la base de N échantillons afin de leur donner une importance équivalente. Cette technique a été appliquée dans de nombreux travaux pour la sélection d'échantillons dans le cadre de développement d'étalonnages multivariés. Par exemple, ZHU *et al.* [116] l'ont appliqué pour le développement de modèles de prédiction de propriétés du bois chinois par spectroscopie visible-PIR à partir de la méthode *Least Square Suport Vector Machine* (LS-SVM). Cette méthode a également été utilisée pour le développement de modèles qualitatifs pour le dépistage de cancer du sein par résonance magnétique nucléaire [46]. Enfin, DE LIRA *et al.* [28] ont employé cette technique pour le contrôle des paramètres de stabilité des biodiesels par spectroscopie MIR.

A.3 Régression PLS

La régression PLS (Partial Least squares) a initialement été développée pour résoudre des problèmes de sciences économiques [77]. Les premières applications de cette méthode

dans le cadre de développement d'étalonnage multivarié datent des années 1980 [71]. L'objectif de la régression PLS est d'établir une équation d'étalonnage linéaire entre la matrice des spectres ($\mathbf{X}$) et la matrice des valeurs de référence ($\mathbf{y}$). Nous pouvons noter que lorsque l'analyse quantitative de plusieurs propriétés doit être réalisée, il est possible d'effectuer une régression PLS sur chacune de ces propriétés (méthode PLS1) ou plusieurs propriétés en même temps (méthode PLS2) [77]. Ici, nous présenterons uniquement la méthode PLS1 (qui sera nommé PLS) car la méthode PLS2 n'a pas été appliquée. L'algorithme PLS décrit ici est celui proposé dans le livre de MARTENS et NÆS [77].

Etape 1 : La matrice spectrale $\mathbf{X}$ et le vecteur des concentrations $\mathbf{y}$ sont centrés pour obtenir les variables $\mathbf{X}_0$ et $\mathbf{y}_0$.

Etape 2 : Pour chaque facteur a= 1, ..., A, les étapes (2.1 à 2.5) sont effectuées pour réaliser l'étalonnage.

Etape 2.1 : Déterminer le vecteur $\mathbf{w}_a$ qui respecte les conditions suivantes :

- Les $\hat{\mathbf{w}}_a$ sont orthogonaux deux à deux : $\hat{\mathbf{w}}_{a-1}^T \cdot \hat{\mathbf{w}}_a = 0$
- $\hat{\mathbf{w}}_a$ est un vecteur unitaire : $\hat{\mathbf{w}}_a^T \cdot \hat{\mathbf{w}}_a$
- $\mathbf{X}_{a-1} = \mathbf{y}_{a-1} \cdot \hat{\mathbf{w}}_a$

On en déduit $\hat{\mathbf{w}}_a = c \cdot \mathbf{X}_{a-1}^T \cdot \mathbf{y}_{a-1}$ où c est un vecteur d'échelle de façon à obtenir le vecteur unitaire $\hat{\mathbf{w}}_a$: $c = \left(\mathbf{y}_{a-1}^T \cdot \mathbf{X}_{a-1} \cdot \mathbf{X}_{a-1}^T \cdot \mathbf{y}_{a-1}\right)$.

Etape 2.2 : Construire les coordonnées factorielles (scores) $\hat{\mathbf{t}}_a$ par projection de $\mathbf{X}_{a-1}$ sur $\hat{\mathbf{w}}_a$.

Il faut résoudre $\mathbf{X}_{a-1} = \hat{\mathbf{t}}_a \cdot \hat{\mathbf{w}}_a^T$

Soit $\hat{\mathbf{t}}_a = \mathbf{X}_{a-1} \cdot \hat{\mathbf{w}}_a \cdot \left(\hat{\mathbf{w}}_a^T \cdot \hat{\mathbf{w}}_a\right)$

Etape 2.3 : Effectuer une régression de $\mathbf{X}_{a-1}$ sur $\hat{\mathbf{t}}_a$ pour obtenir le vecteur propre spectral $\hat{\mathbf{p}}_a$

L'équation $\mathbf{X}_{a-1} = \hat{\mathbf{t}}_a \cdot \hat{\mathbf{p}}_a^T$

a pour solution $\hat{\mathbf{p}}_a = \mathbf{X}_{a-1}^T \cdot \hat{\mathbf{t}}_a \cdot \left(\hat{\mathbf{t}}_a^T \cdot \hat{\mathbf{t}}_a\right)$

Etape 2.4 : Calculer la valeur propre de la réponse $\hat{q}_a$

en déterminant la solution de $\mathbf{y}_{a-1} = \hat{\mathbf{t}}_a \cdot \hat{q}_a$
qui s'écrit $\hat{q}_a = \mathbf{y}_{a-1}^T \cdot \hat{\mathbf{t}}_a \cdot \left(\hat{\mathbf{t}}_a^T \cdot \hat{\mathbf{t}}_a\right)$

Etape 2.5 : Soustraire la contribution de la composante a à la matrice spectrale $\mathbf{X}$ et au vecteur de réponse $\mathbf{y}$. Cette étape est appelée *deflation* en anglais :
$\mathbf{X}_a = \mathbf{X}_{a-1} - \hat{\mathbf{t}}_a \cdot \hat{\mathbf{p}}_a^T$
$\mathbf{y}_a = \mathbf{y}_{a-1} - \hat{\mathbf{t}}_a \cdot \hat{q}_a$

Etape 3 : Calculer les coefficients PLS qui correspondent à l'équation d'étalonnage :
La solution de l'équation $\mathbf{y} = \mathbf{X} \cdot \hat{\beta}$
est $\hat{\beta} = \hat{\mathbf{W}} \cdot \left(\hat{\mathbf{P}}^T \cdot \hat{\mathbf{W}}\right)^{-1} \cdot \hat{\mathbf{q}}$
où $\hat{\mathbf{W}}, \hat{\mathbf{P}}$ et $\hat{\mathbf{q}}$ contiennent respectivement les $\hat{\mathbf{w}}_{a=1\ldots A}, \hat{\mathbf{p}}_{a=1\ldots A}$ et $\hat{q}_{a=1\ldots A}$.

Etape 4 : L'équation d'étalonnage est ensuite utilisée pour prédire la valeur d'un échantillon inconnu
$\mathbf{y}_{inconnu} = \mathbf{X}_{inconnu} \cdot \hat{\beta}$

La régression PLS comporte l'avantage de gérer les cas où les variables sont très corrélées et que le nombre de variables (longueurs d'onde) est beaucoup plus grand que le nombre d'observations (nombre d'échantillons) [100]. Elle permet également d'interpréter les relations entre les variables prédictives et la réponse par le biais des vecteurs propres des variables ou des coefficients PLS. Le seul point critique de la régression PLS est le choix du nombre de facteurs. Généralement, ce nombre de facteurs est déterminé par validation croisée qui est l'objet de la partie suivante.

A.4 Validation croisée

La validation croisée est une méthode pour l'estimation des performances d'un modèle. Comme évoqué dans la partie précédente, le point critique dans le cadre des méthodes de régression sur facteurs est le choix du nombre de facteurs. En effet, la prise en compte d'un nombre de facteurs trop important tend à ajouter du bruit dans l'équation d'étalonnage et provoque un sur-ajustement [85]. Or, l'erreur d'étalonnage diminue généralement

avec l'ajout de facteurs tandis que l'erreur de prédiction admet un minima (Figure A.2). Afin de fixer ce nombre de facteur optimal, il est d'usage d'utiliser la validation croisée. Tout comme le *bootstrap* (Partie 3.5.2), la validation croisée est une méthode de ré-échantillonnage. En effet, en validation croisée, les échantillons d'étalonnage servent à la fois à l'élaboration du modèle et à l'évaluation de ses performances. En effet, le principe est de retirer un échantillon (validation croisée totale) ou un groupe d'échantillons (validation croisée partielle), de construire un modèle avec les échantillons restants et de prédire les échantillons écartés en utilisant l'équation d'étalonnage ainsi déterminée. Dans le cas de la validation croisée totale, cette opération est répétée pour chaque échantillon et pour un nombre de facteurs compris entre 1 et une valeur maximale définie par l'utilisateur. En validation croisée partielle, il faut définir le taille du groupe d'échantillons à retirer, si le tirage est effectué avec ou sans remise et le nombre de groupes retirés pour chaque nombre de facteurs.

Quelque soit la méthode appliquée, la racine carrée de la moyenne des erreurs quadratiques de validation croisée (RMSECV) est calculée pour chaque composante. Lors de la validation croisée, l'ajout de facteurs va diminuer la valeur de RMSECV jusqu'à une valeur minimale qui correspond idéalement au nombre de facteurs optimal devant correspondre à l'erreur de prédiction la plus faible (Figure A.2).

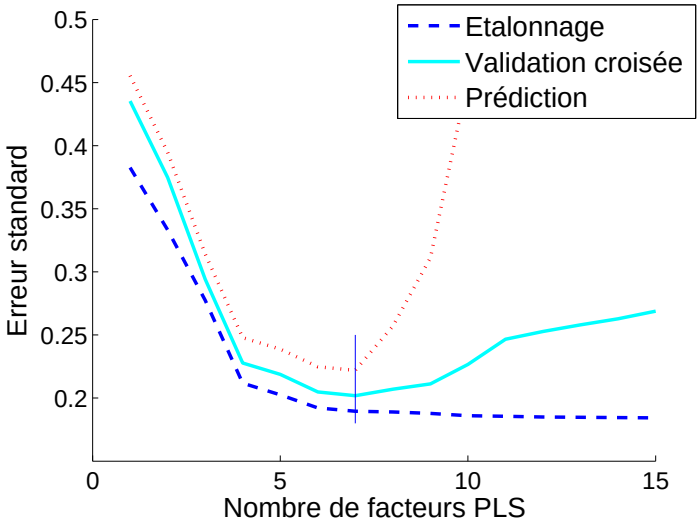


FIGURE A.2 – Choix du nombre de composantes par validation croisée [77]

A.5 Critères statistiques

Les critères statistiques qui sont calculés pour évaluer les performances d'un modèle prédictif sont rappelés ici, à savoir, le pourcentage de variance expliquée, le coefficient de détermination, le biais et la racine carrée de la moyenne des erreurs quadratiques.

Le pourcentage de variance expliquée correspond à la variance expliquée par le modèle par rapport à la variance totale. Elle peut se calculer sur les descripteurs ($\mathbf{X}$) ou sur la réponse ($\mathbf{y}$) et pour chaque composante. La variance expliquée sur $\mathbf{X}$ est parfois utilisée pour fixer le nombre de composantes. Ce choix est cependant discutable car la variance expliquée ne prend pas en compte les corrélations entre $\mathbf{X}$ et $\mathbf{y}$. Ainsi, un facteur qui explique une faible variance sur $\mathbf{X}$ mais très corrélée à $\mathbf{y}$ peut améliorer de manière significative le modèle.

Le coefficient de détermination est un indicateur qui permet de juger la qualité d'une régression. Il mesure l'adéquation entre le modèle et les données observées. Sa valeur est comprise entre entre 0 et 1. Il est défini comme la part de variance expliquée par rapport à la variance totale. Tout comme le pourcentage de variance expliquée, il peut se calculer sur $\mathbf{X}$ (R^2) ou sur $\mathbf{y}$ (Q^2).

Le biais est défini comme une erreur systématique dans une évaluation statistique.

Dans le cadre d'un modèle prédictif, le biais sert donc à estimer s'il existe une erreur systématique entre la valeur prédite $\hat{y}_i$ et la valeur de référence y_i. Il se calcule suivant l'équation A.6.

$$Biais = \frac{\sum_{j=1}^{n} (y_i - \hat{y}_i)}{n} \tag{A.6}$$

Le biais peut être calculé sur le lot d'étalonnage ou sur le lot de validation. Cependant, le biais calculé sur le lot d'étalonnage est très faible car les méthodes de régression ont pour but de minimiser les résidus.

Enfin, la racine carrée de la moyenne des erreurs quadratiques ou *Root Mean Square Error* (RMSE) est calculée pour estimer l'erreur de prédiction sur un lot d'échantillons :

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{A.7}$$

L'erreur de RMSE peut être calculée :

- sur les échantillons du lot d'étalonnage (RMSEC)
- lors de la validation croisée (RMSECV)
- sur les échantillons du lot de validation (RMSEP)

B

Résultats complémentaires

Les résultats présentés ici complètent l'étude de comparaison des spectroscopies MIR et PIR et la présentation des modèles de prédiction de l'ensemble des propriétés des produits lourds.

B.1 Comparaison des spectroscopies MIR et PIR

Nous présenterons ici les résultats obtenus lors de la comparaison des spectroscopies MIR et PIR pour la détermination de la densité, des teneurs en hydrogène, en carbones insaturés et en carbone Conradson. La démarche mise en œuvre pour ces propriétés est identique à celle présentée dans la partie 4.3. Pour chaque propriété, le Tableau B.1 indique le prétraitement appliqué, le domaine spectral considéré et le nombre de facteurs PLS utilisé. Les erreur de RMSEP obtenues par les modèles développés à partir des spectres MIR et PIR sont également spécifiées. Enfin, la valeur de significativité *p-value* calculée à l'aide du test "randomisation t-test" est également indiquée.

Pour les mêmes raisons qu'évoquées dans la Partie 4.3, les modèles calculés à partir des spectres MIR sont développés en appliquant la dérivée 1$^{\text{ère}}$ sur les domaines spectraux 3200-2700 cm^{-1} et 1700-650 cm^{-1}. Lorsque les spectres PIR sont utilisés, les modèles de prédiction de la densité, de la teneur en hydrogène et de la teneur en carbones insaturés sont développés sur le domaine 7000-4000 cm^{-1} en appliquant la méthode *Weighted Least Square baseline* avec un polynôme d'ordre 3. En effet, ces paramètres permettent d'éliminer les variations de la ligne de base qui, d'après les résultats obtenus, ne correspondent

pas une information pertinente.

En revanche, pour la prédiction de la teneur en carbone Conradson, l'erreur de RMSEP la plus faible a été obtenue sur les spectres PIR dérivés et sur le domaine 9000-4000 cm^{-1}. L'analyse du carbone Conradson consiste à peser le résidu de produit après chauffage de l'échantillon à forte température (500°C) pendant 15 minutes. Les molécules qui composent le résidu sont les molécules dont les points d'ébullition sont les plus hauts. La plupart de ces molécules appartiennent donc à la fraction asphaltène. La remontée de la ligne de base est donc une information utile pour la détermination de la teneur en carbone Conradson.

Pour la prédiction de la teneur en carbones insaturés, les modèles développés à partir des spectroscopie MIR et PIR amènent à des RMSEP de 1,46 %(m/m) et 1,18 %(m/m), respectivement. Avec une valeur de *p-value* inférieure à 0,05, le test de comparaison a démontré que l'erreur de RMSEP du modèle PIR est significativement plus faible que celle du modèle MIR. En revanche, le test de comparaison n'indique pas de différence significative pour la détermination des autres propriétés. En effet, les valeurs de RMSEP obtenues par les modèles développés à partir des spectres MIR et PIR sont très proches : respectivement, 0,0041 et 0,0035 pour la densité, 0,18 %(m/m) et 0,17 pour la teneur en hydrogène et 1,11 %(m/m)pour les deux modèles de prédiction de la teneur en carbone Conradson.

Tableau B.1 – Présentation et comparaison des modèles PLS, MB-PLS et S-PLS pour la prédiction des teneurs en saturés, aromatiques, résines et asphaltènes C7

Propriétés	**Spectroscopie**	**domaine spectral** (cm^{-1})	**Prétraitement**	**Nombre de facteurs PLS**	**RMSEP**	***p-value***
Densité	MIR	3200-2700 1700-650	Dérivée 1ère	5	0,0041	0,23
	PIR	7000-4000	WLSB(3)[a]	6	0,0035	
Hydrogène	MIR	3200-2700 1700-650	Dérivée 1ère	5	0,18 %(m/m)	0,40
	PIR	7000-4000	WLSB(3)[a]	6	0,17 %(m/m)	
Carbones insaturés	MIR	3200-2700 1700-650	Dérivée 1ère	3	1,46 %(m/m)	$\leq$ 0,05
	PIR	7000-4000	WLSB(3)[a]	5	1,18 %(m/m)	
Carbone Conradson	MIR	3200-2700 1700-650	Dérivée 1ère	7	1,11 %(m/m)	0,49
	PIR	9000-4000	Dérivée 1ère	7	1,11 %(m/m)	

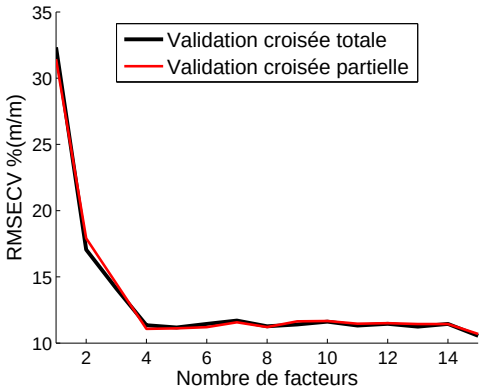
[a] : méthode *Weighted Least Square baseline* avec un polynôme d'ordre 3

B.2 Caractérisation globale des produits lourds par spectroscopie PIR

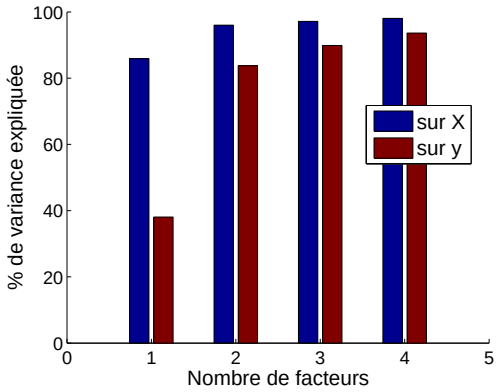
Les modèles de prédiction décrits ici ont été jugés insatisfaisants. Nous les présentons pour illustrer le type de problèmes rencontrés pour le développement de ces étalonnages multivariés.

B.2.1 La viscosité

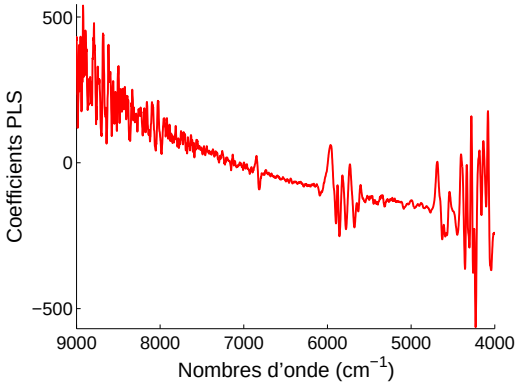
Le modèle de prédiction de la viscosité a été développé à partir de 153 échantillons sur une gamme analytique allant de 3,45 à 212,6 cSt. Nous pouvons noter que l'échantillon extrême (793,19 cSt) qui avait été repéré dans la Partie 3.2.1 a été éliminé. Ce modèle a été développé avec 4 facteurs sur la gamme 9000-4000 cm^{-1} et les spectres ont été analysés en dérivée $1^{ère}$. Nous pouvons remarquer que les coefficients sont très bruités, notamment entre 9000 et 7500 cm^{-1}. Ceci s'explique par le fait que les spectres dont la remontée de la ligne de base est importante sont bruités sur ce domaine spectrale. Le graphique des résidus de prédiction illustre que le modèle obtenu n'est pas acceptable (Figure B.1d). Nous pouvons remarquer d'une part que les faibles valeurs en viscosité sont très mal prédites. D'autre part, la Figure B.1e montre que pour les plus fortes valeurs, un biais très important est observé. Comme nous l'avons évoqué dans la Partie 2.2.2, les produits lourds du pétrole sont des fluides non-newtoniens ce qui entraîne une relation non-linéaire avec la viscosité. Plusieurs modèles ont été testés pour tenter de pallier cette non-linéarité. Nous avons tout d'abord divisé la base en deux groupes représentant les DSV (faibles valeurs de viscosité) et les RA (fortes valeurs de viscosité). En effet, les valeurs de viscosité de ces deux coupes sont déterminées par deux méthodes différentes (la viscosité cinématique pour les DSV et la viscosité dynamique pour les RA). Une conversion des valeurs de viscosité dynamique est ensuite réalisée (Partie 1.3.1). Cette approche n'a cependant pas permis d'obtenir des modèles satisfaisants. Des modèles ont également été calculés en considérant le logarithme des valeurs de référence mais sans résultat.



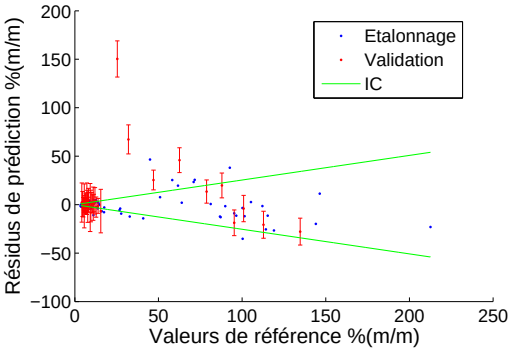
(a) Validation croisée pour le choix du nombre de facteurs PLS



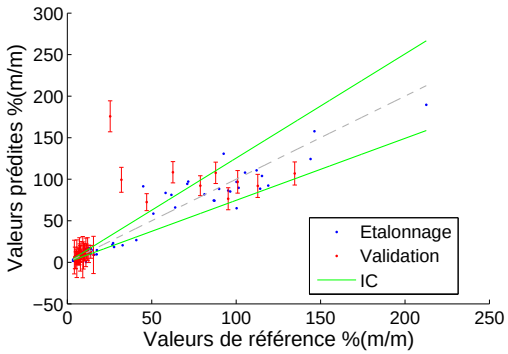
(b) Pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction

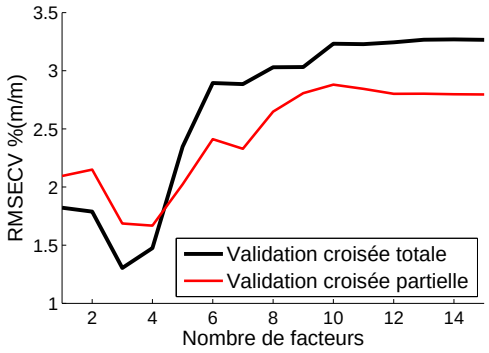


(e) Diagramme de parité

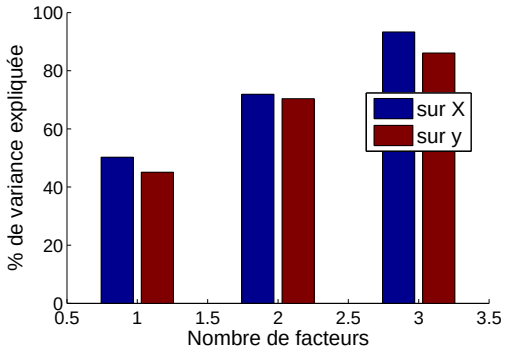
FIGURE B.1 – Interprétation du modèle de prédiction de la teneur en viscosité

B.2.2 La teneur en asphaltènes

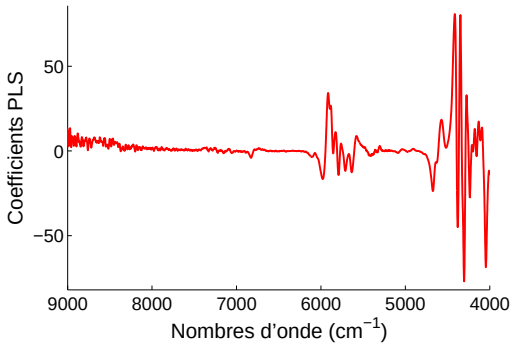
Du fait du faible nombre d'échantillons disponibles pour la teneur en asphaltènes, l'équation d'étalonnage a donc été calculée à partir de 17 échantillons qui couvrait la gamme 3,5 - 12,8 %(m/m). Le lot d'échantillons de validation était composé de seulement 4 échantillons. Le modèle de prédiction a été développé à partir de 3 facteurs PLS, sur le domaine 9000-4000 cm^{-1} des spectres prétraités en dérivée $1^{ère}$. La Figure B.2b illustre que la variance expliquée sur $\mathbf{y}$ est faible. Nous pouvons également constater que l'influence de l'absorption électronique des liaisons $n - \pi^*$ et $\pi - \pi^*$ est faible ce qui est surprenant. L'erreur de prédiction obtenue est relativement faible (RMSEP = 0,85%). Nous pouvons cependant remarquer qu'un biais est présent sur les prédictions des échantillons du lot de validation (Figure B.2e). C'est pourquoi ce modèle a été jugé insatisfaisant.



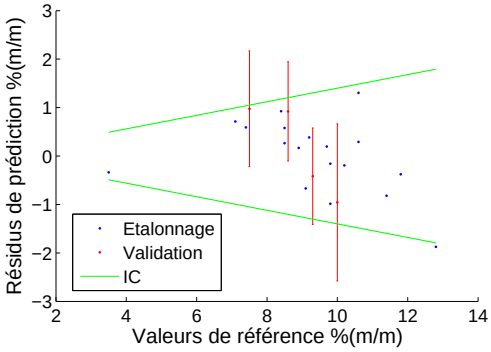
(a) Validation croisée pour le choix du nombre de facteurs PLS



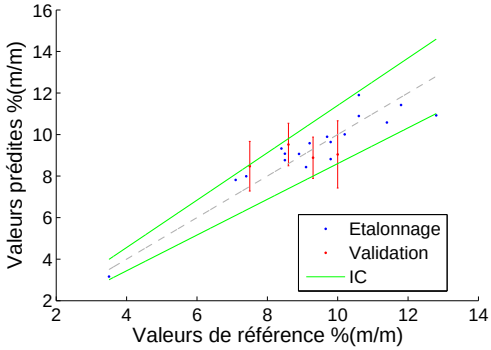
(b) Pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction

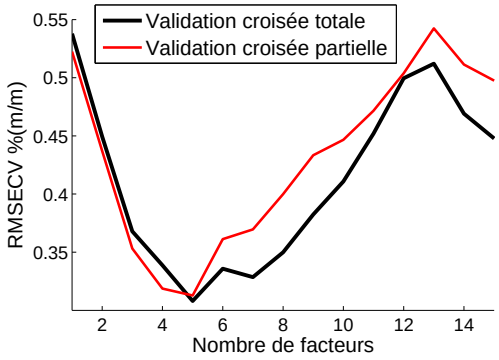


(e) Diagramme de parité

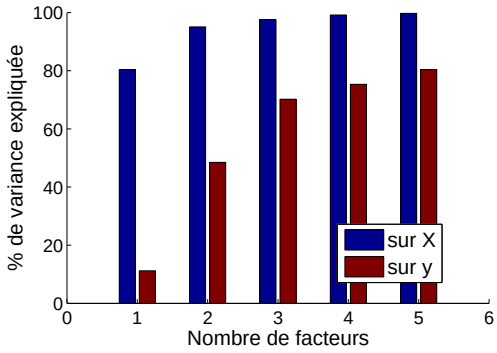
FIGURE B.2 – Résultats pour le modèle de détermination de la teneur en asphaltènes

B.2.3 La teneur en carbone

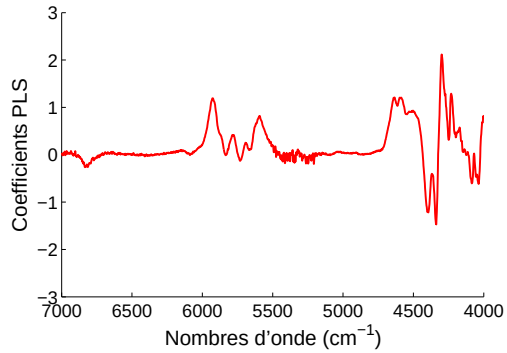
Le nombre d'échantillons disponibles pour la teneur en carbone est faible (41 échantillon). Un modèle a tout de même été développé sur la gamme 86,01 - 88,40 %(m/m) pour une étude de faisabilité. Pour cela, le domaine 7000-4000 cm^{-1} a été utilisé avec un prétraitement WLSB. La Figure B.3a montre que l'erreur de RMSECV admet un minimum pour 5 facteurs PLS. Le pourcentage cumulé de variance expliquée sur $\mathbf{y}$ est néanmoins très faible pour 5 facteurs ($< 80\%$). La valeur de RMSEP obtenue est faible (0,5%). Cependant, les Figures B.3d et B.3e illustrent qu'il existe un biais très important dans ce modèle. Ce modèle n'est donc pas satisfaisant. A partir de cette étude, il est difficile de conclure sur la faisabilité de la détermination de la teneur en carbone. La collecte de nouveaux échantillons apparaît donc indispensable.



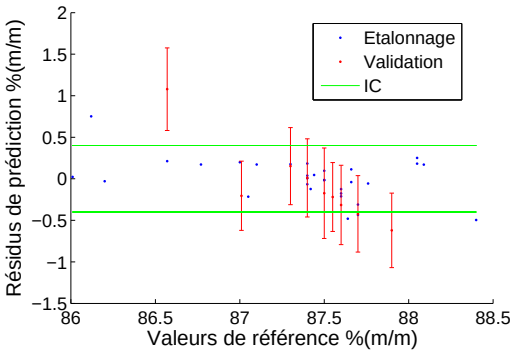
(a) Validation croisée pour le choix du nombre de facteurs PLS



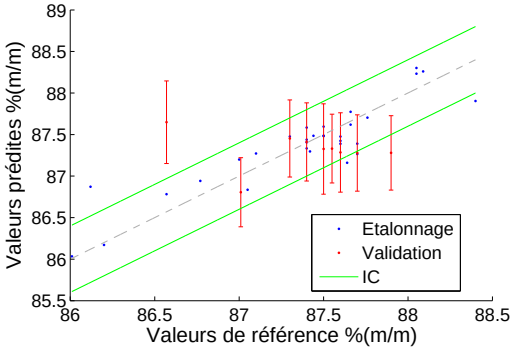
(b) Pourcentage cumulé de variance expliquée sur $\mathbf{X}$ et sur $\mathbf{y}$ en fonction du nombre de facteurs PLS



(c) Coefficients de l'équation d'étalonnage



(d) Résidus de prédiction



(e) Diagramme de parité

FIGURE B.3 – Résultats pour le modèle de détermination de la teneur en carbone

C

Communications Scientifiques sur la période de la thèse

Publication

Jérémy Laxalde, Cyril Ruckebusch, Olivier Devos, Noémie Caillol, François Wahl, Ludovic Duponchel : Characterisation of Heavy Oils using Near-Infrared Spectroscopy : Optimisation of Pre-processing Methods and Variable Selection, *Analytica Chimica Acta*, 705(1-2) :227-234 ; 2011

Communication orale

Jérémy Laxalde, Cyril Ruckebusch, Noémie Caillol, François Wahl, Ludovic Duponchel : Heavy Oil Characterisation by Vibrational Spectroscopy : Comparison of NIR and MIR – Evaluation of Combination Methods, *15th International Conference on Near Infrared Spectroscopy – Cape Town – Afrique du Sud* , 2011

Communication par affiche

Jérémy Laxalde, Cyril Ruckebusch, Olivier Devos, Noémie Caillol, François Wahl, Ludovic Duponchel : Characterisation of Heavy Oils using Near Infrared (NIR) Spectroscopy and Chemometrics, *CAC Conference 2010 (Chemometrics in Analytical Chemistry) – Anvers – Belgique*, 2010

ANALYSE DES PRODUITS LOURDS DU PETROLE PAR SPECTROSCOPIE INFRAROUGE

Résumé : L'objectif de cette thèse est le développement d'une analyse rapide pour la caractérisation des produits lourds du pétrole. Des modèles de prédiction de propriétés des produits lourds ont été développés à partir des spectroscopies moyen infrarouge (MIR) et proche infrarouge (PIR). Ce travail a principalement porté sur l'optimisation des modèles prédictifs des teneurs en composés saturés, aromatiques, résines et asphaltènes (SARA). Une optimisation simultanée par algorithmes génétiques du choix des prétraitements des données spectrales et des variables à sélectionner a été évaluée. Cette approche a permis de conduire au meilleur pouvoir prédictif des modèles PIR et a montré le potentiel d'interprétation des variables sélectionnées. Une étude de comparaison des modèles développés séparément à partir des spectres MIR et PIR a ensuite été réalisée. La spectroscopie PIR s'est révélée être globalement plus performante dans le cadre de notre application. Il a également été démontré que la fusion de données spectroscopiques pouvait améliorer la qualité des prédictions. Au vu des résultats, il semble nécessaire que les modèles développés séparément à partir de ces spectroscopies conduisent à des performances similaires pour espérer une amélioration lors de la fusion des données spectrales. Le potentiel de l'interprétation des techniques de régression à blocs multiples a également été confirmé pour identifier les informations spectrales spécifiques contenues dans les spectres MIR et PIR. Enfin, les modèles de prédiction de la densité, des teneurs en SARA, en carbone Conradson, en hydrogène, en soufre et en azote ont été jugés satisfaisants pour une utilisation au laboratoire.

Mots Clés : produits lourds du pétrole ; spectroscopie proche infrarouge ; spectroscopie moyen infrarouge ; chimiométrie ; algorithmes génétiques ; méthodes *multiblock* ;

ANALYSIS OF HEAVY OIL PRODUCTS BY INFRARED SPECTROSCOPY

Abstract: The aim of this study is to develop an alternative analysis for the characterisation of heavy oil products. Predictive chemometric models have been developed by mid-infrared (MIR) and near infrared (NIR) spectroscopies. This work is mainly concerned with the predictive model optimisation of saturate, aromatic, resin and asphalten contents (SARA). A simultaneous optimisation procedure of spectral data pre-processing methods and variable selection by genetic algorithms was evaluated. This approach led to the best NIR predictions and showed the potential interpretation of the selected variables. A comparative study of MIR and NIR spectroscopies for the development of heavy oil property predictive model was also performed. Results have shown that NIR spectroscopy is globally better for our application. It has been shown that spectroscopic data fusion can improve predictive power of models. It seems however necessary that both spectroscopies, when considered separately, correspond to similar predictive power in order to expect an improvement when combining MIR and NIR. The interpretation potential of multiblock has been confirmed for the identification of MIR and NIR specific information. Finally, models developed for the prediction of density, contents of SARA, Conradson carbon, hydrogen, sulphur and nitrogen were found satisfactory for an application at laboratory.

Keywords: heavy oils products; near-infrared spectroscopy; mid-infrared spectroscopy; chemometrics; genetic algorithms; multiblock methods;